Real time updating of the flood frequency distribution through data assimilation

Cristina Aguilar[1], Alberto Montanari[2], María José Polo[1]

[1]Fluvial dynamics and hydrology research group, Andalusian Institute of Earth System Research, University of Córdoba, Cordoba, 14071, Spain
[2]Department DICAM, University of Bologna, Bologna, 40136, Italy

Correspondence to: Cristina Aguilar (caguilar@uco.es, formerly: caguilar@ugr.es)

Abstract

We explore the memory properties of catchments for predicting the likelihood of floods basing on observations of average flows in pre-flood seasons. Our approach assumes that flood formation is driven by the superimposition of short and long term perturbations. The former is given by the short term meteorological forcing leading to infiltration and/or saturation excess, while the latter is originated by higher-than-usual storage in the catchment. To exploit the above sensitivity to long term perturbations, a Meta-Gaussian model and a data assimilation approach is implemented for updating the flood frequency distribution a season in advance. Accordingly, the peak flow in the flood season is predicted in probabilistic terms by exploiting its dependence on the average flow in the antecedent seasons. We focus on the Po River at Pontelagoscuro and the Danube river at Bratislava. We found that the shape of the flood frequency distribution is noticeably impacted by higher-than-usual flows occurring up to several months earlier. The proposed technique may allow one to reduce the uncertainty associated to the estimation of flood frequency.

1 Introduction

The physical, chemical and ecological state of processes leading to the formation and quality of river flow is characterized by persistence at several different time scales (Koutsoyiannis, 2014). In fact, anomalous conditions for such processes, such as those generated by extreme meteorological events, may produce a long-lasting impact on the river flow, depending on climatic and catchment behaviors (Lo and Famiglietti, 2010). For instance, flood generation is impacted by the initial soil moisture

condition of the catchment, which may be in turn impacted by groundwater levels that are related to global catchment storage (Massari et al., 2014). Persistence can be exploited to improve river flow forecasting at seasonal-to-interannual time scale. Furthermore, persistence provides useful indications to better understand the functioning of a catchment and the dynamics of the water cycle.

5 Indeed, the study of persistence is one of the most classical research endeavors in hydrology, since the early works by Rippl (1883) and Hazen (1914) on the estimation of the optimal storage for reservoirs. Hurst (1951) investigated the Nile River flows while working at the design of the Aswan Dam and postulated that geophysical records may be affected by a complex form of persistence that may last for a long time (O' Connell et al., 2015). Later on, Thomas and Fiering (1962) and Yevjevich (1963)

10 introduced autoregressive models for annual and seasonal streamflow simulation therefore stimulating the development of subsequent models of increasing complexity for simulating hydrological persistence.

Recently, the attention has been focused on long term persistence (LTP), which is associated with the Hurst-Kolmogorov behavior (Koutsoyiannis, 2011). LTP manifests itself through a power-law decay of

15 the autocorrelation function of the process, which implies that the summation of the autocorrelation coefficients diverges to infinity (Montanari et al., 1997). LTP implies the possible presence of long term cycles (Beran, 1994), which in turn means that perturbations of hydrological processes may last for a long time, therefore providing a possible explanation for the occurrence of clusters of extreme hydrological events, such as floods and droughts (Montanari, 2012). LTP also has implications in the

20 study of climate change, as it is connected with an enhanced natural variability of climatic processes (Koutsoyiannis and Montanari, 2007).

While LTP has been long studied, limited attempts have been made to exploit LTP in data assimilation procedures for improving streamflow forecasting. The motivation probably is that LTP is recognized to exert a noticeable impact on the river flow volume over long time scales, while its effect on the

25 magnitude of single events is less noticeable. Nevertheless, the presence of LTP and seasonal correlation necessarily affects flood frequency, to an extent that has been poorly explored.

The present contribution aims to enhance our understanding of the persistence properties of river flows to improve seasonal river flow forecasting. By taking inspiration from the idea that the probability of

extreme floods may be increased by long term stress, like higher than usual rainfall lasting for several months, the research question that we address here can be stated as follows: Can higher than usual river discharges in the previous season be associated to a higher probability of floods in the subsequent high flow season? The quantification of the effect of antecedent flows for different time lags on the occurrence of floods would help to assess how long a river remembers its past (Aguilar et al., 2016). From a technical point of view, we aim to propose a technique for updating a season in advance the flood frequency distribution estimated for a given river, through a data assimilation approach, by exploiting the information provided by river flows in the pre-flood seasons.

It is interesting to highlight that the state of a catchment, and in particular its storage, is affected by previous precipitation. Therefore, it would be reasonable to exploit the information provided by previous rainfall rather than previous flows for the sake of updating the flood frequency distribution. However, areal rainfall estimation for catchments with large extension and complex orography is affected by large uncertainty (Moulin et al., 2009). Therefore, we utilize here flows during pre-flood seasons as a proxy for catchment storage instead of rainfall. While the above assumption may be reasonable, one should consider that it may not hold when the river flows are impacted by massive regulation.

2 Study sites and data sources

We focus our attention on two large basins, namely, the Po river basin at Pontelagoscuro (Italy) and the Danube river basin at Bratislava (Slovakia). The Po River is the longest river entirely flowing in the Italian Peninsula (Fig. 1) with a catchment area of about 71000 km^2 at the delta. The average annual precipitation in the catchment is 78 km^3 in volume, of which 60% reaches the closure river cross-section at Pontelagoscuro. The hydrological behavior of the Po River is described in detail in recent studies (Zanchettin et al., 2008; Montanari, 2012; Zampieri et al., 2015). The discharge pattern at Pontelagoscuro presents a mean annual flow of about 1470 m^3s^{-1} and shows a typical pluvial regime, and thus a strong seasonality with two flood seasons in spring and autumn (Fig. 2). An intense exploitation of water resources for irrigation, hydro-power production, civil and industrial use is found

3

in the catchment. Even though water resources management is currently sustainable on average, critical situations are experienced during drought periods (Montanari, 2012).

The upper Danube basin drains from the northern side of the Alps and the southern area of the central European Highlands into Bratislava in a 131331 km^2 catchment area where the mean annual flow is about 2053 m^{3}s^{-1}. The hydrological behavior of the upper Danube basin can be found in detail in the literature (Nester et al., 2011; Blöschl et al., 2013). The average annual precipitation in the catchment is 123 km^3 and the discharge pattern shows a typical alpine regime and thus a strong seasonality with one flood season in the summer (Fig. 2).

Daily discharge and monthly precipitation and temperature data for the Po and Danube river basins were analyzed in this study. The observation periods as well as descriptive statistics of the different time series are shown in Table 1. Discharge time series at Pontelagoscuro for the Po River and Bratislava for the Danube River were provided, respectively, by the Regional Agency for Environmental Protection (ARPA)— Emilia Romagna, Hydro-meteorological Office and by Global Runoff Data Center (GRDC, 2011). The series are not affected by missing values. They correspond to a time span of 90 and 107 years for Po and Danube, respectively.

The Po river is regulated by the presence of several dams as reservoirs for hydroelectricity production, which are mainly located in the Alpine region. Also, the outflow from the lakes Como, Garda, Iseo, Idro and Maggiore is regulated (Zanchettin et al., 2008). These regulations do not noticeably impact the trend and the low-frequency variability of the peak flows, while they may affect the low flows at daily and sub-daily time scale (Zampieri et al., 2015). The upper part of the Danube has been ideal for building hydropower plants and up to 59 dams are found along the river's first 1000 km. As stated in the Danube River Basin Management Plan, stretches in the very upper part of the river may present noticeably altered flows. (Maps 7a, b, c in DRBM, 2009). The effect of regulation on peak flows in Slovakia is deemed to be negligible, while low and average flows may be noticeably impacted.

Precipitation and temperature time series were calculated based on weather data sets obtained from the HISTALP project (Auer et al., 2007). Only weather stations where sufficiently long data sets are available were used (Table 1). The study period was conditioned by the availability of discharge data

even though both meteorological variables were available for a longer historical period. For each study site, catchment area average precipitation and temperature time series were constructed using Thiessen polygons.

3 Methodology

5 In order to address the research question outlined in Section 1, namely, to verify the opportunity of updating the flood frequency distribution a season in advance by exploiting the information provided by the river flow in a given pre-flood season, we perform an analysis of the memory properties of the hydrological cycle in the considered catchments. We first focus on meteorological variables, namely, temperature and mean areal rainfall to check whether a memory pattern is detectable in the weather.

10 Rainfall and temperature are considered as they are the main drivers of river flow, with temperature being particularly influential on the lower values. Then, we turn to the direct analysis of river flows. We first estimate the Hurst exponent (H) for the considered time series, to verify whether the hypothesis of the presence of LTP is supported by data evidence. Then, we turn to the analysis of the statistical dependence between the peak flow in the flood season and the average flow during the previous season,

15 to empirically check whether updating the flood frequency distribution produces useful results. Results from the latter analysis are assessed in view of the LTP estimation.

3.1 Estimation of long term persistence

Assessment of long term persistence for hydrological data has been presented by several contributions (see, for instance, Szolgayova et al., 2014; and Zampieri et al., 2015 for analyses carried out for the

20 river flows of the Danube and Po River, respectively). Time series with long-term memory or persistence exhibit a power-law decay of the autocorrelation function (Beran, 1994), that is:

$$\rho(k) \sim c_k \cdot k^{2H-2} \quad k \to \infty \tag{1}$$

where $\rho(k)$ is the autocorrelation function of the process at lag k, c_k is a constant and $H \in [0\ 1]$ is the Hurst exponent or the intensity of the LTP (Montanari et al., 1997). For a stationary process, H is

constrained in the range [0.5,1). A value equal to 0.5 means absence of LTP; the higher the *H*, the higher the intensity of LTP.

In this work, *H* was estimated by using different heuristic methods. In detail, we applied the rescaled range (R/S) analysis, the aggregated variance method (climacogram; see Dimitriadis and Koutsyiannis, 2015), and the differenced variance method. An extended description of numerous methodologies to assess the persistence properties of time series to provide support to the possible presence of the Hurst-Kolmogorov behavior can be found in Taqqu et al. (1995), Montanari et al. (1996, 1997, 2000) and Koutsoyiannis (2003).

A strong seasonal component in the different hydrological variables in both study time series has been reported by the literature (e.g. Montanari, 2012; Szolgayova et al., 2014; Zampieri et al, 2015). It is well known that a strong seasonality often implies the presence of periodic deterministic components in the data that can introduce a bias in LTP estimation (Montanari et al., 1997, 2000). Also, the presence of slowly decaying or increasing trends may induce a bias as well. Thus, prior to long-term memory assessments, all time series were detrended and deseasonalized. For each time series, a 366-term (for daily data) and 13-term (monthly data) moving average for a trend approximation was applied, followed by a stable seasonal filter for removing of the seasonal cycle (Brockwell et al., 2002).

3.2 Analysis of the peak flow dependence on average flows during pre-flood seasons

In order to analyze the stochastic connection between the average river flows in the antecedent seasons and the average and peak flow in the flood season, a bivariate probability distribution function was fitted. In what follows, random variables and their outcomes are identified with bold and un-bold characters, respectively. The yearly variables analyzed in this study were:

-The monthly mean flow in the given pre-flood season (independent or explanatory variable), $\boldsymbol{Q_m}$.

-The peak flow in the flood season or annual maximum daily flow (dependent variable), $\boldsymbol{Q_p}$.

-The mean daily flow in the flood season (dependent variable), $\boldsymbol{Q_{mf}}$.

A meta-Gaussian model (Kelly and Krzysztofowicz, 1997; Montanari and Brath, 2004) is used to model the joint probability distribution between the selected explanatory and dependent variables. The method involves the following steps.

First, the time series $Q_m(t)$, $Q_p(t)$ and $Q_{mf}(t)$ with sample size n, where n is the number of years in the observation period, are extracted from the observed datasets. Then, the Normal Quantile Transform (NQT) is applied in order to make their marginal probability distributions Gaussian, therefore obtaining the normalized observations $NQ_m(t)$ and $NQ_p(t)$ and $NQ_{mf}(t)$.

The NQT is a non parametric transformation that can be applied to normalize any arbitrarily distributed random variable. There are numerous applications of the NQT in hydrological studies, to generate flow samples from specified marginal distributions (Moran, 1970; Hosking and Wallis, 1988), to perform Bayesian updating of prior distributions (Kelly and Krzysztofowicz, 1994), to model bivariate distributions with arbitrary marginal distribution (Krzysztofowicz et al., 1994; Aguilar et al., 2016). The NQT is adopted within the Bayesian Forecasting System for river flows (Krzysztofowicz and Kelly, 2000; Krzysztofowicz and Herr, 2001; Krzysztofowicz and Maranzano 2004a, b; Maranzano and Krzysztofowicz 2004). It was also applied for assessing the uncertainty of rainfall-runoff simulations (Montanari and Brath, 2004; Montanari and Grossi, 2008; Bogner et al., 2012), to deseasonalise hydrological time series (Montanari, 2005). Being free of any distributional assumption, the NQT allows one to avoid the selection of a suitable parametric model for the distribution of the considered hydrological variable.

The NQT involves the following steps when we take $\boldsymbol{Q_m}$ as an example: (1) Sorting the sample of $Q_m(t)$ from the smallest to the largest observation, Q_{m1}, ..., Q_{mn}; (2) estimating the cumulative frequency FQ_{mi} by using the Weibull plotting position (Stedinger et al., 1993); (3) for each FQ_{mi} the standard normal quantile NQ_{mi} is computed as $NQ_{mi} = G^{-1}(FQ_{mi})$, with G denoting the standard normal distribution and G^{-1} its inverse, and associated with the corresponding Q_{mi}. Thus, a discrete mapping of Q_{mi} to its transformed counterpart NQ_{mi} is obtained. In order to apply the inverse of the NQT for any NQ_{mi}, linear interpolation is applied to connect the points of the discrete mapping previously obtained. Bogner et al. (2012) propose different parametric and non-parametric approaches for the extrapolation of extreme

values. In this study, the region beyond the maximum and the minimum available NQ_{mi} values is covered by linear extrapolation.

Finally, the meta-Gaussian model (Kelly and Krzysztofowicz, 1997; Montanari and Brath, 2004) is fitted between the random explanatory variable and each random dependent variable in their canonical form in the Gaussian domain. In what follows, we specify the equations for the peak flow as the dependent variable. We assume: (1) stationarity and ergodicity of both $\mathbf{NQ_m}$ and $\mathbf{NQ_p}$; and (2) that the cross dependence between both $\mathbf{NQ_m}$ and $\mathbf{NQ_p}$ can be represented by the normal linear equation:

$$NQ_p(t) = \rho(\mathbf{NQ_m}, \mathbf{NQ_p}) \cdot NQ_m(t) + N\varepsilon(t) \tag{2}$$

where $\rho(\mathbf{NQ_m}, \mathbf{NQ_p})$ is the Pearson's cross-correlation coefficient between $\mathbf{NQ_m}$ and $\mathbf{NQ_p}$, and $N\varepsilon$ is an outcome of the stochastic process $\mathbf{N\Theta}$, which is independent, homoscedastic, stochastically independent of $\mathbf{NQ_m}$ and normally distributed with zero mean and variance $1-\rho^2(\mathbf{NQ_m}, \mathbf{NQ_p})$. The parameters of the bivariate probability distribution function are the mean ($\mu(\mathbf{NQ_m})$=0 and $\mu(\mathbf{NQ_p})$=0), the standard deviation ($\sigma(\mathbf{NQ_m})$=1 and $\sigma(\mathbf{NQ_p})$=1) of the normalized series, and the Pearson's cross-correlation coefficient between both normalized series, $\rho(\mathbf{NQ_m}, \mathbf{NQ_p})$. In the presence of dependence between $\mathbf{NQ_m}$ and $\mathbf{NQ_p}$, the correlation coefficient will be significantly different from zero. The bivariate Gaussian distribution implies that, for an arbitrary (observed) $NQ_m(t)$, the probability distribution function of $\mathbf{NQ_p}$ is Gaussian, with parameters (Eq. 3 and 4):

$$\mu(\mathbf{NQ_p}) = \rho(\mathbf{NQ_m}, \mathbf{NQ_p}) \cdot NQ_m(t) \tag{3}$$

$$\sigma(\mathbf{NQ_p}) = \left(1 - \rho^2(\mathbf{NQ_m}, \mathbf{NQ_p})\right)^{0.5} \tag{4}$$

Then, by taking the inverse of the NQT one can infer the updated probability distribution of $\mathbf{Q_p}$ conditioned to the observed outcome $Q_m(t)$.

In order to verify the validity of the linear model (Eq. 2), an evaluation based on the behavior of the residuals is applied. Following the graphical approach proposed by Cook and Weisberg (1994), the residual plot of $N\varepsilon(t)$ versus $\rho(\mathbf{NQ_m}, \mathbf{NQ_p}) \cdot NQ_m(t)$ should not show any systematic trend under the

target model. Curve trends or fan shape trends indicate non-linear cross dependence and variability of the variance of $\mathbf{N\Theta}$, respectively (Montanari and Brath, 2004).

The same methodology was applied for the other dependent variable considered in this study, $\boldsymbol{Q_{mf}}$. Therefore, once the parameters of each distribution are computed, the probability distribution function of both the peak flow and the mean flow in the flood season can be updated after observing the mean flow in the considered pre-flood season.

The proposed methodology involves uncertainty in the estimated flood frequency distributions which is mainly given by two sources: the first is uncertainty in the NQT, namely, uncertainty in the estimation of the marginal probability distribution of independent and dependent variables in the regression. The second source of uncertainty is related to the estimation of the cross-correlation coefficient between dependent and independent variables in the Gaussian domain. The NQT is a non-parametric transformation and therefore its uncertainty cannot be determined quantitatively (Maranzano and Krzytofowicz, 2004; Montanari and Brath, 2004). To reduce uncertainty, it is advisable that NQT is estimated by using long records encompassing a wide range of meteorological and hydrological conditions. Uncertainty in the cross-correlation coefficient can be quantified for a given confidence level and again depends on the length of the records. A quantitative estimation of uncertainty for the cross-correlation coefficient was carried out in both study sites. Uncertainty bounds at the 95% confidence level are computed by first computing the Fisher's transformation,

$$z\left(NQ_m, NQ_p\right) = 0.5 \cdot ln\left(\left(1 + \rho\left(NQ_m, NQ_p\right)\right) \cdot \left(1 - \rho\left(NQ_m, NQ_p\right)\right)^{-1}\right) \tag{5}$$

where the random variable $\mathbf{z}$ is approximately normally distributed with a standard deviation of:

$$\sigma(\mathbf{z}) = ((n\text{-}3)^{-1})^{3^{-1}} \tag{6}$$

Therefore, confidence bands for $z\left(NQ_m, NQ_p\right)$ can be computed at a given confidence level which can be converted to the confidence bands for $\rho(\mathbf{NQ_m}, \mathbf{NQ_p})$ by taking Fisher' s inverse transformation. If a negative (positive) value for the lower (upper) confidence limit is obtained for a positive (negative)

9

estimated value of $\rho(NQ_m, NQ_p)$ then we reset the lower (upper) limit to 0. Finally, the limiting flood frequency distributions can be obtained for the lower and upper value of $\rho(NQ_m, NQ_p)$.

In order to infer the actual impact of the dependence between peak flows and mean flow in the flood season with the mean flow in the pre-flood seasons, the unconditioned flood frequency distribution and the updated distributions inferred for several higher-than-average values of mean flow (e.g. 70%, 80% and 95% quantiles) in a given pre-flood season were compared. We assume that peak flows can be adequately modeled through the Extreme Value Type 1 (EV1) distribution and we present a comparison between the unconditioned peak flows frequency distribution and the updated peak flows frequency distributions.

Finally, a leave-one-out validation analysis was carried out to emulate a real-world application. We removed from the analysis the data observed in the year with the wettest pre-flood season (1977 in the Po and 1944 in the Danube) and then we estimated the probability distribution for the peak flow in the flood season for that year. Uncertainty was estimated for this application.

3.2.1 Identification of the flood season

According to previous studies in the literature, directional statistics (Mardia, 1972) represents an effective method for identifying the timing of hydrological extreme events (e.g. Castellarin et al., 2001; Cunderlik and Burn, 2002; Baratti et al., 2012). Following Bayliss and Jones (1993), the date of occurrence of an event i (e.g. maximum annual daily flow) can be transformed into a directional statistic by converting the Julian date of occurrence, J_{di}, into an angular measure, θ_i, through Eq. (7):

$$\theta_i = J_{di} \cdot (2\pi \cdot 365^{-1}) \tag{7}$$

Each date of occurrence can then be written in polar coordinates by means of a vector with a unit magnitude and the direction specified by Eq. (7). Therefore, the x_p and y_p coordinates of the mean of the sample of n dates of occurrence can be computed with Eq. (8):

$$x_p = n^{-1} \cdot \sum_{i=1}^{n} \cos(\theta_i) \qquad y_p = n^{-1} \cdot \sum_{i=1}^{n} \sin(\theta_i) \tag{8}$$

The direction, $\bar{\theta}$, and magnitude, r, of the mean in polar coordinates can then be obtained by Eq. (9) and Eq. (10) respectively. Equation (9) gives a measure of the mean timing of the event for the sample of dates, and can be converted back to a mean Julian date, M_D, through Eq. (7). Equation (10) indicates the regularity or seasonality of the phenomenon. Values of r close to one imply a strong regularity in the

5   dates of occurrence of the event considered. In contrast, values of r close to zero indicate a great dispersion and thus, a great inter-annual variability in the dates of occurrence of the event throughout the year.

$$\bar{\theta} = \arctan\left(y_{\mathrm{p}} \cdot x_p^{-1}\right) \tag{9}$$

$$r = (x^2 + y^2)^{-2} \tag{10}$$

10   Finally, the limits of the occurrence of the phenomenon can quantitatively be identified by adding and subtracting to $\bar{\theta}$, the standard deviation in radians, σ, given by Eq. (11):

$$\sigma = \left(-2 \cdot \ln(r)\right)^{-2} \tag{11}$$

We applied directional statistics to the following variables in order to identify the flood season in each study site: 1) annual maximum series of daily flows (AMD); 2) high flow events defined from

15   frequency analysis as those events when the daily discharge exceeds the 95th percentile, Q_{95}, for longer than 15 days. Results are shown in a circle plot where each date of occurrence of the variables analyzed in the data set are visible along the perimeter. The month of occurrence of each of the variables can be easily identified. Also, the proximity to the center of the circle of the global value indicates the regularity of the phenomenon with the highest regularity found in the perimeter of the circle.

20   **4 Results and discussion**

4.1 Long term persistence estimation

The application of the heuristic methods for LTP estimation to deseasonalized and detrended time series is displayed in Table 2. H values above 0.5 were obtained for the mean daily river flows in both rivers and thus, all three heuristic methods detect the presence of noticeable LTP. The intensity of LTP seems

25   to be more or less the same for monthly flow data. Similarly, H values in monthly temperature data of

0.64 and 0.61 in Po and Danube, respectively, suggest the presence of LTP in both records. In contrast, the estimated *H* values in the monthly rainfall datasets are not sensibly higher than 0.5.

In general, these results agree with previous outcomes of long term persistence studies for the daily discharge of the Po at Pontelagoscuro (Montanari, 2012) as well as with previous studies in the daily river flows in an upstream tributary of the Po (*H*=0.71-0.81) and in the monthly rainfall registered at certain weather stations within the watershed (Montanari et al., 1996; 1997). Also, *H* values of the same order of magnitude were found by Szolgayova et al. (2014) for the rainfall (*H*=0.43-0.50) and temperature (*H*=0.65-0.72) monthly time series in the upper Danube watershed at Bratislava.

4.2 Meta-Gaussian model for updating the flood frequency distribution

4.2.1 Flood season identification

Figure 3 shows the results of the directional statistics applied to the extreme events in both rivers. In the Po river, we can see a very low regularity (*r*≈0.1) and high dispersion (4 months) in the annual maximum daily flows (AMD in Fig. 3) due to their possible occurrence in any of the two high flow seasons, spring and autumn, as depicted in Fig. 2. The seasonality increases to *r* values close to 0.8 for high flow events that mostly take place in autumn as already reported in previous studies (Zanchettin et al., 2008; Montanari, 2012).

In the Danube, we find a considerable regularity in high flow events (*r*≈0.8) but a certain decrease in the annual maximum flows (with *r* values of 0.4). Nevertheless, the 2-month dispersion in the date of occurrence is lower than in the Po river and corresponds to the length of the high flow season reported in Fig.2. In view of these results we set October-November and May-July as the main flood seasons in the Po and Danube respectively.

As pre-flood season, we consider a 1-month period, which is long enough in order to reduce the effect of river regulation. We first set the month preceding the flood season (i.e., September and April for Po and Danube, respectively) as pre-flood season. Then, we repeat the analysis by making reference to the previous months, with the expectation that the statistical dependence decreases as the pre-flood season is moved back in the past.

12

4.2.2 Estimation of the Meta-Gaussian model

Table 3 shows the cross-correlation coefficients $\rho(\mathbf{NQ_m,NQ_p})$ and $\rho(\mathbf{NQ_m,NQ_{mf}})$, along with their confidence bands, between the normalized dependent variables ($\mathbf{NQ_p}$ and $\mathbf{NQ_{mf}}$ in both study sites) and the explanatory variable ($\mathbf{NQ_m}$) at each study site. In detail, we assumed that $\boldsymbol{Q_m}$ is given by the monthly mean flow in each of the 9 months preceding the floods season (from September to January in the Po river and from April to August in the antecedent year in the upper Danube). Table 3 shows that the correlation coefficient decreases as the considered pre-flood season moves backwards, as we expected. Besides, we always found noticeably higher coefficients with the mean flow in the flood season ($\rho(\mathbf{NQ_m, NQ_{mf}})$), than with the annual maximum daily flows ($\rho(\mathbf{NQ_m, NQ_p})$) in both rivers. For example, a cross-correlation coefficient of 0.24 was obtained between $\mathbf{NQ_p}$ and $\mathbf{NQ_m}$ in the Po when the pre-flood season considered is September, compared to 0.39 between $\mathbf{NQ_{mf}}$ and the same explanatory variable, $\mathbf{NQ_m}$. Moreover, a continuous decreasing cross-correlation coefficient is found as we move further apart the flood season and negative correlation in the Po river appears from May-June backwards. These negative correlations put in evidence that low flows in the winter season may be related to higher flows in the summer season and therefore higher peak flows in the fall season. The latter outcome could be explained by a higher storage during the winter months in the form of increased snowpack, which may be related to the frequency and memory properties of temperature and precipitation data.

The only anomalous correlation is found when considering the $\boldsymbol{Q_m}$ in March as the explanatory variable for both dependent variables in the Danube. This month corresponds with both the peak in the snowmelt annual cycle in the catchment (Zampieri et al., 2015) and the steepest rising slope in the hydrograph (Fig. 2). Therefore, the use of monthly mean flow might not be representative given the high variability in the daily flows along this month and the complexity of the processes that are affecting the streamflow (complex contribution from subsurface flow or from the runoff generated from snowmelt/precipitation). An evaluation was carried out for the Meta-Gaussian model by using residuals plots (Montanari and Brath, 2004). Fig. 4 shows the residuals for a time span of 4-months backwards the flood season at each study site. The residuals look homoscedastic, therefore confirming that the model assumptions about the residuals behavior are justified.

4.2.3 Flood frequency distribution updating

In order to decipher the technical benefit that can be gained by updating the flood frequency distribution through the proposed data assimilation procedure, we assumed that above average river flows are observed in the month preceding the flood season and then applied the Meta-Gaussian model to estimate the updated probability distribution. In detail, we assume that, in average, monthly flow corresponding to the 70%, 80% and 95% quantile is observed in September for the Po River and April for the Danube River.

Fig. 5 and 6 show the unconditioned and updated probability density functions (pdf) of the normalized peak flow (i.e., the peak flow transformed to the canonical Gaussian distribution). As one would expect, the results show that the higher the cross-correlation value, the lower the variability in the distribution of the normalized dependent variable and the higher the mean value. For example, in the Po River for the occurrence of the 95th quantile value in the normalized mean flow in September, the pdf is centered around a mean vaue of 0.4 and presents a standard deviation of 0.97 (Fig. 5). In contrast, if one attempts to estimate the probability distribution of NQ_p conditioned to the occurrence of the 95th quantile of the normalized mean flow in July, no noticeable change is found in the estimate with respect to the unconditioned distribution. In fact, the resulting probability density function (pdf) for NQ_p is centered around a mean value of 0.09 with a standard deviation of 0.998. The same behavior is found in the probability distribution of the other dependent variable in its normalized form, NQ_{mf}, where the higher correlation coefficients (Table 3) determine even a greater displacement with respect to the unconditioned distribution. In fact, the pdf of NQ_{mf} conditioned to the occurrence of the 95th quantile value in the normalized mean flow in September is centered around a mean vaue of 0.64 and presents a standard deviation of 0.92 (Fig. 5).

In the upper Danube a similar scheme is found with the mean of the probability distribution of NQ_p and NQ_{mf} conditioned to the occurrence of the 95th quantile of the normalized mean flow in April, displaced to 0.32 and 0.82 respectively (Fig. 6 and Table 3).

Figure 7 shows the comparison between the unconditioned flood frequency distribution and the updated distributions in the untransformed domain when the flow in the previous month (September for the Po

14

River, April for the Danube River) is higher than usual (70%, 80% and 95% quantile). For example, in the Po river, the unconditioned flood for a return period of 200 years, which results equal to 12507 m^3s^{-1}, increases up to 13790 m^3s^{-1} (about 10% increase) when the mean flow in September corresponds to its 95% quantile. Similarly, in the upper Danube the unconditioned peak flow for a return period of

5   200 years, 10075 m^3s^{-1}, increases up to 10861 m^3s^{-1} (about 8% increase) when the mean flow in April corresponds to its 95% quantile. The differences show that the average flow during the pre-flood seasons may indeed provide useful indications to update the flood frequency distribution.

After removing from the analysis the observations of the years 1977 for the Po River and 1944 for the Danube River, which are the previous flood season wettest years in record, an emulation of a one-

10   month-ahead real time prediction of the probability distribution of the flood flows in the next flood season was developed, along with uncertainty estimation as described in Section 3.2. Figure 8 shows, for both Po and Danube rivers, the unconditioned flood frequency distribution along with the updated one. 95% confidence bands for the latter are also shown. It can be seen that the proposed procedure allows one to obtain an effective update in real world applications.

15

5 Conclusions

The analysis of the observed mean daily flow values suggests the existence of LTP in both study sites with H values above 0.71. Such persistence is exploited to improve streamflow forecasting in the flood season in terms of the mean monthly flow of the pre-flood seasons. To this end, we automatically detect

20   the flood season through directional statistics and we fit a bivariate Gaussian distribution function to model the above dependence. A 10% and 8% increase in the 200-yr return period peak flows are found in the Po and Danube, respectively, when the average flows during the previous month corresponds to its 95% quantile. The above results show that the Meta-Gaussian model applied to the streamflow records can be used for updating a season in advance the flood frequency distribution estimated for a

25   given river, through a data assimilation approach by using the mean monthly flow of the pre-flood seasons.

15

The methodology herein proposed can be applied to any other study site once the parameters of the Meta-Gaussian model confirm the presence of the above stochastic dependence. Like in any time series analysis method, records that encompass a wide range of meteorological and hydrological conditions should be used to minimize uncertainty, which is in this case related to the estimation of the correlation coefficient and standardization of the regression variables. Finally, other explanatory variables (e.g. rainfall, snowmelt, etc.) can be incorporated to profit from additional stochastic dependence among peak flows and the state of the catchment and external forcings.

The findings presented in this paper highlight that river memory has an impact on flood formation and should then be properly considered for real time management of flood risk mitigation and resilience of societal settings to floods. The procedure herein described can provide useful information in those cases where the memory of the catchment is supposed to persist for a long time. These conditions may occur when the precipitation-runoff transformation is characterized by a slow development. Memory is frequently found to be related to the storage capacity of the catchment and the complexity of the river network. Therefore, they may be indicators of potential useful results from the proposed approach.
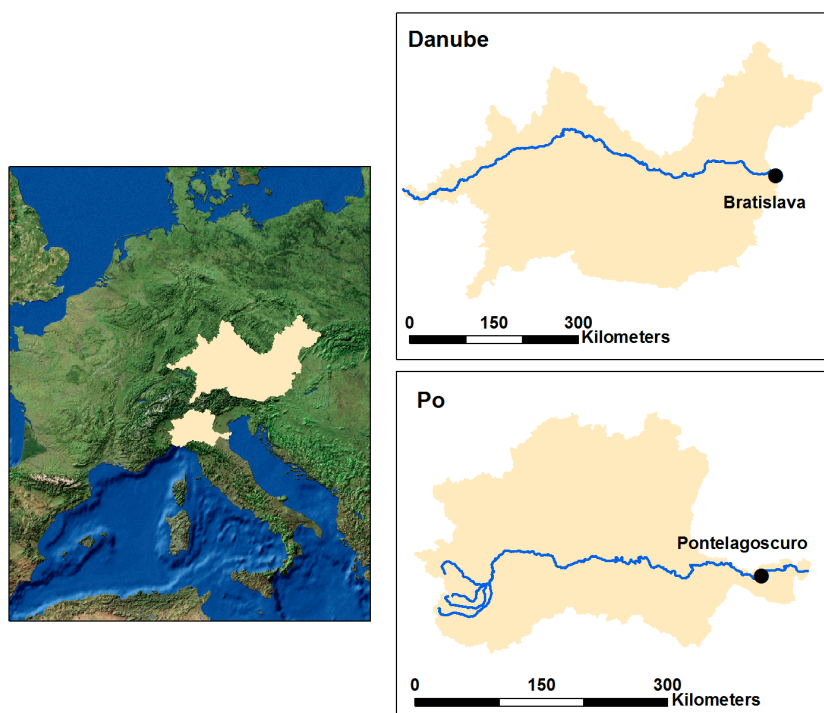
Acknowledgements

References

Aguilar, C., Montanari, A., and Polo, M. J., Long term prediction of flood occurrence, Proc. IAHS, 373, 189-192, doi:10.5194/piahs-373-189-2016, 2016.

Auer, I., Böhm, R., Jurkovic, A., Lipa, W., Orlik, A., Potzmann, R., Schöner, W., Ungersböck, M., Matulla, C., Briffa, K., Jones, P., Efthymiadis, D., Brunetti, M., Nanni, T., Maugeri, M., Mercalli, L., Mestre, O., Moisselin, J.M., Begert, M., Müller- Westermeier, G., Kveton, V., Bochnicek, O., Stastny, P., Lapin, M., Szalai, S., Szentimrey, T., Cegnar, T., Dolinar, M., Gajic-Capka, M., Zaninovic, K., Majstorovic, Z. and Nieplova, E.: HISTALP – historical instrumental climatological surface time series of the Greater Alpine Region, Int. J. Climatol. 27, 17-46, doi: 10.1002/joc.1377, 2007.

Baratti, E., Montanari, A., Castellarin, A., Salinas, J.L., Viglione, A. and Brezzi, A.: Estimating the flood frequency distribution at seasonal and annual time scales, Hydrol. Earth Syst. Sci., 16, 4651-4660, doi:10.5194/hess-16-4651-2012, 2012.

Bayliss, A. C. and Jones, R. C.: Peaks-over-threshold flood database: summary statistics and seasonality, Crowmarsh Gif- ford, Rep. 121, Institute of Hydrology, Wallingford, 61 pp., 1993.

Beran, J.: Statistics for long-memory processes, Chapman & Hall/CRC, New York, 1994.

Blöschl, G., Nester, T., Komma, J., Parajka, J. and Perdigão, R. A. P.: The June 2013 flood in the Upper Danube Basin, and comparisons with the 2002, 1954 and 1899 floods, Hydrol. Earth Syst. Sci., 17, 5197-5212, doi:10.5194/hess-17-5197-2013, 2013.

Bogner, K., Pappenberger, F. and Cloke, H. L.: Technical Note: The normal quantile transformation and its application in a flood forecasting system, Hydrol. Earth Syst. Sci., 16, 1085-1094, doi: 10.5194/hess-16-1085-2012, 2012.

Brockwell, P. J. and Davis, R. A.: Introduction to Time Series and Forecasting, Vol. 1, Taylor & Francis, 2002.

Castellarin, A., Burn, D. H. and Brath, A.: Assessing the effectiveness of hydrological similarity measures for flood frequency analysis, J. Hydrol., 241, 270-285, doi:10.1016/S0022-1694(00)00383-8, 2001.

Cook, R. D. and Weisberg, S.: An Introduction to Regression Graphics, John Wiley, Hoboken, N. J., 1994.

Cunderlik, J. and Burn, D. H.: Analysis of the linkage between rain and flood regime and its application to regional flood frequency estimation, J. Hydrol., 261, 115–131, 2002.

Dimitriadis, P. and Koutsoyiannis, D.: Climacogram versus autocovariance and power spectrum in stochastic modeling for Markovian and Hurst–Kolmogorov processes. Stoch. Environ. Res. Risk Assess., 29, 1649-1669, doi:10.1007/s00477-015-1023-7, 2015.

GRDC. Global Runoff Data Center − Dataset of River Discharge Time Series. Klobenz, Germany, 2011.

Hazen, A.: Storage to be provided in impounding municipal water supply, Trans. Amer. Soc. Civil Eng., 77(1), 1539-1640, 1914.

Hosking, J. R. M. and Wallis, J.R.: The effect of intersite dependence on regional flood frequency analysis, Water Resour. Res., 24, 588-600, doi: 10.1029/WR024i004p00588, 1988.

Hurst, H. E.: Long-term storage capacity of reservoirs, Trans. Amer. Soc. Civil Eng., 116, 770-808, 1951.

Kelly, K. S. and Krzysztofowicz, R.: Probability distributions for flood warning systems, Water Resour. Res., 30(4), 1145-1152, 1994.

Kelly, K. S. and Krzysztofowicz, R.: A bivariate meta-Gaussian density for use in hydrology, Stochastic Hydrol. Hydraul., 11, 17–31, DOI: 10.1007/BF02428423, 1997.

Koutsoyiannis, D.: Climate change, the Hurst phenomenon, and hydrological statistics, Hydrol. Sci. J., 48, 3-24, 2003.

Koutsoyiannis, D.: Hurst-Kolmogorov dynamics and uncertainty, J. Amer. Water Resour. Assoc., 47 (3), 481-495, doi: 10.1111/j.1752-1688.2011.00543.x, 2011.

Koutsoyiannis, D.: Entropy: from thermodynamics to hydrology, Entropy, 16 (3), 1287-1314, doi: 10.3390/e16031287, 2014.

Koutsoyiannis, D. and Montanari, A.: Statistical analysis of hydroclimatic time series: Uncertainty and insights, Water Resour. Res., 43 (5), W05429, doi:10.1029/2006WR005592, 2007.

Krzysztofowicz, R.: Transformation and normalization of variates with specified distributions, J. Hydrol., 197, 286–292, 1997.

Krzysztofowicz, R. and Herr, H. D.: Hydrologic uncertainty processor for probabilistic river stage forecasting: Precipitation-dependent model, J. Hydrol., 249, 46-68, 2001.

Krzysztofowicz, R. and Kelly, K. S.: Hydrologic uncertainty processor for probabilistic river stage forecasting, Water Resour. Res., 36, 3265-3277, doi: 10.1029/2000WR900108, 2000.

5   Krzysztofowicz, R. and Maranzano, C. J.: Hydrologic uncertainty processor for probabilistic stage transition forecasting, J. Hydrol., 293, 57-73, 2004a.

Krzysztofowicz, R. and Maranzano, C.J.: Bayesian system for probabilistic stage transition forecasting, J. Hydrol., 299, 15-44, 2004b.

Krzysztofowicz, R., Kelly, K.S. and Long, D.: Reliability of flood warning systems, J. Water Resour.
10   Plann. Manage., 120(6), 906-926, 1994.

Lo, M. H. and Famiglietti, J. S.: Effect of water table dynamics on land surface hydrologic memory, J. Geophys. Res., 115, D22118, doi:10.1029/2010JD014191, 2010.

Mardia, K. V.: Statistics of Directional Data, Academic, San Diego, Calif., 1972.

Massari, C., Brocca, L., Moramarco, T., Tramblay, Y. and Lescot, J. F. D.: Potential of soil moisture
15   observations in flood modelling: estimating initial conditions and correcting rainfall, Adv. Water Resour., 74, 44-53, 2014.

Maranzano, C.J. and Krzysztofowicz, R.: Identification of likelihood and prior dependence structures for hydrologic uncertainty processor, J. Hydrol., 290, 1-21, 2004.

Montanari, A.: Deseasonalisation of hydrological time series through the normal quantile transform, J.
20   Hydrol., 313, 274-282, doi: 10.1016/j.jhydrol.2005.03.008, 2005.

Montanari, A.: Hydrology of the Po River: looking for changing patterns in river discharge, Hydrol. Earth Syst. Sci., 16, 3739-3747, doi: 10.5194/hess-16-3739-2012, 2012.

Montanari, A. and Brath, A.: A stochastic approach for assessing the uncertainty of rainfall-runoff simulations, Water Resour. Res., 40, W01106, doi:10.1029/2003WR002540, 2004.

25   Montanari, A. and Grossi, G.: Estimating the uncertainty of hydrological forecasts: A statistical approach, Water Resour. Res., 44, W00B08, doi:10.1029/2008WR006897, 2008.

Montanari, A., Rosso, R., and Taqqu, M. S.: Some long-run properties of rainfall records in Italy, J. Geophys. Res., 101: 431-438, doi: 0148-0227/96/96JD-0251, 1996.

19

Montanari, A., Rosso, R., and Taqqu, M. S.: Fractionally differenced ARIMA models applied to hydrologic time series: identification, estimation and simulation, Water Resour. Res., 33, 1035-1044, 1997.

Montanari, A., Rosso, R., and Taqqu, M. S.: A seasonal fractional ARIMA model applied to the Nile
5   river monthly flows at Aswan, Water Resour. Res., 36, 1249-1259, 2000.

Moran, P.: Simulation and Evaluation of Complex Water Systems Operations, Water Resour. Res., 6, 1737-1742, doi: 10.1029/WR006i006p01737, 1970.

Moulin, L., Gaume, E. and Obled, C.: Uncertainties on mean areal precipitation: assessment and impact on streamflow simulations. Hydrol. Earth Syst. Sci., 13(2), 99-114, 2009.

10   Nester, T., Kirnbauer, R., Gutknecht, D., and Blöschl, G.: Climate and catchment controls on the performance of regional flood simulations, J. Hydrol., 402, 340-356, 2011.

Rippl, W.: The capacity of storage-reservoirs for water-supply, in: Minutes of the Proceedings of the Institution of Civil Engineers, Vol. 71, No. 1883, pp. 270-278, Thomas Telford-ICE Virtual Library, 1883.

15   Stedinger, J. R., Vogel, R. M. and Foufoula-Georgiou, E.: Frequency analysis of extreme events, in Handbook of Hydrology, edited by D. R. Maidment, pp. 18.1-18.66, McGraw-Hill, New York, 1993.

Szolgayova, E., Parajka, J., Blöschl, G., and Bucher, C.: Long term variability of the Danube River flow and its relation to precipitation and air temperature, J. Hydrol., 519: 871-880, doi: 10.1016/j.hydrol.2014.07.047, 2014.

20   Taqqu, M. S., Teverovski, V., and Willinger, W.: Estimators for long-range dependence: an empirical study, Fractals, 3, 785–798, 1995.

Thomas, H. A. and Fiering, M. B.: Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation, Design of water resource systems, 459-493, 1962.

Yevjevich, V. M.: Fluctuations of wet and dry years, I, Research data assembly and mathematical
25   models, Hydrol. Pap. I, 1963.

Zampieri, M. Scoccimarro, E., Gualdi, S. and Navarra, A.: Observed shift towards earlier spring discharge in the main Alpine rivers, Sci. Tot. Envir., 503-504, 222-232, doi: 10.1016/j.scitotenv.2014.06.036, 2015.

Zanchettin, D., Traverso, P. and Tomasino, M.: Po River discharge: a preliminary analysis of a 200-year time series, Climatic Change, 88, 411-433, doi: 10.1007/s10584-008-9395-z, 2008.

5

Figure 1: Study sites. Danube river basin at Bratislava and Po river basin at Pontelagoscuro.
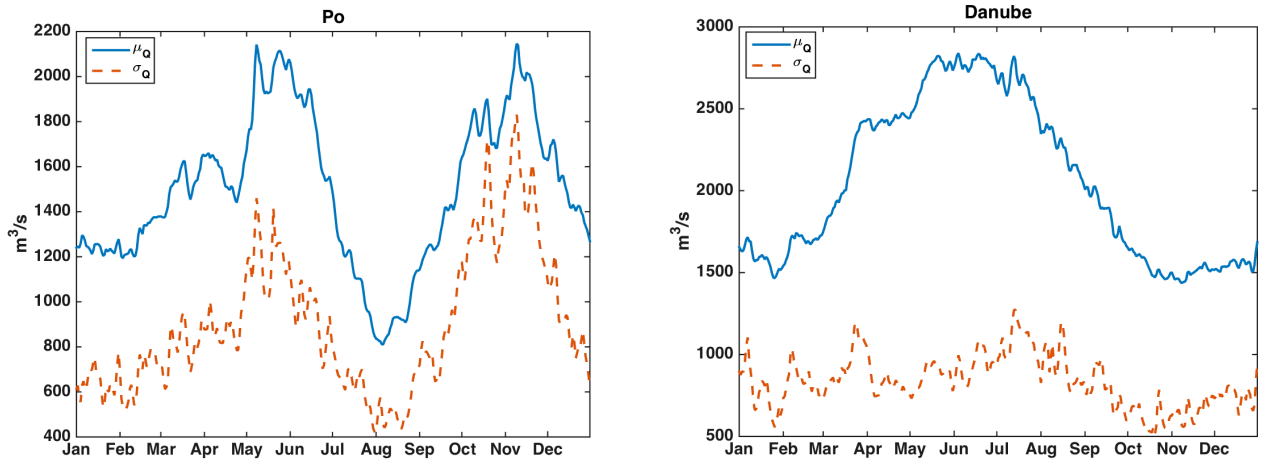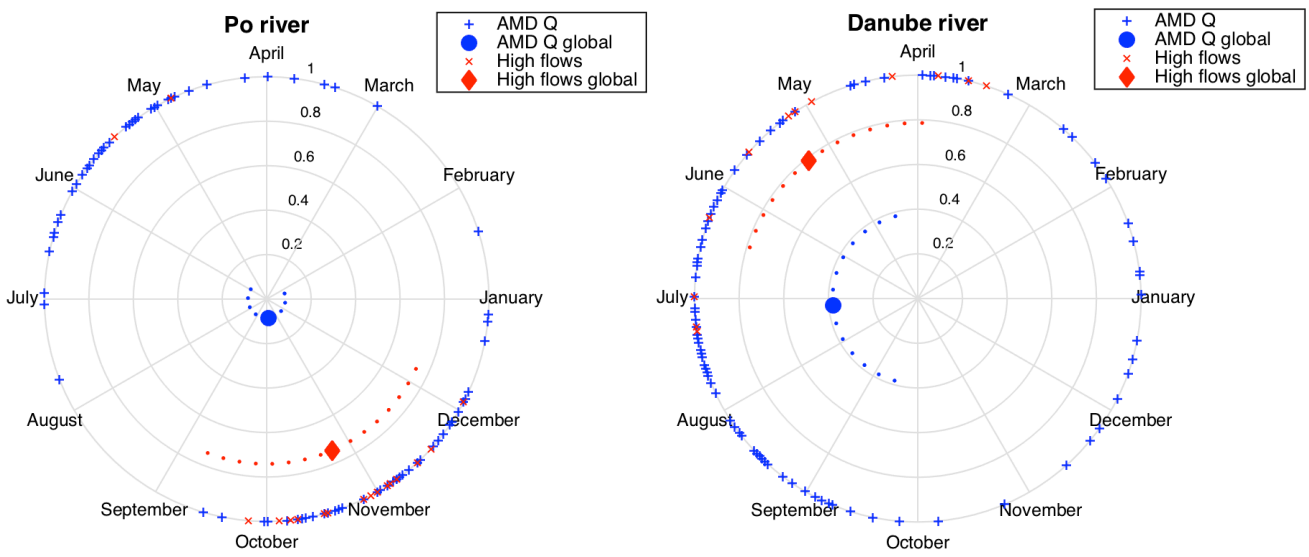
Figure 2: Daily mean value μ_Q (m^{3}s^{-1}) and daily standard deviation σ_Q (m^{3}s^{-1}) of the daily flows in the observation periods: 1920-2009 in the Po at Pontelagoscuro, 1901-2007 in the Danube at Bratislava

5

Figure 3. Seasonality space representation of the annual maximum daily flows (AMD) and high flow events. Dots around the global value indicate the dispersion.
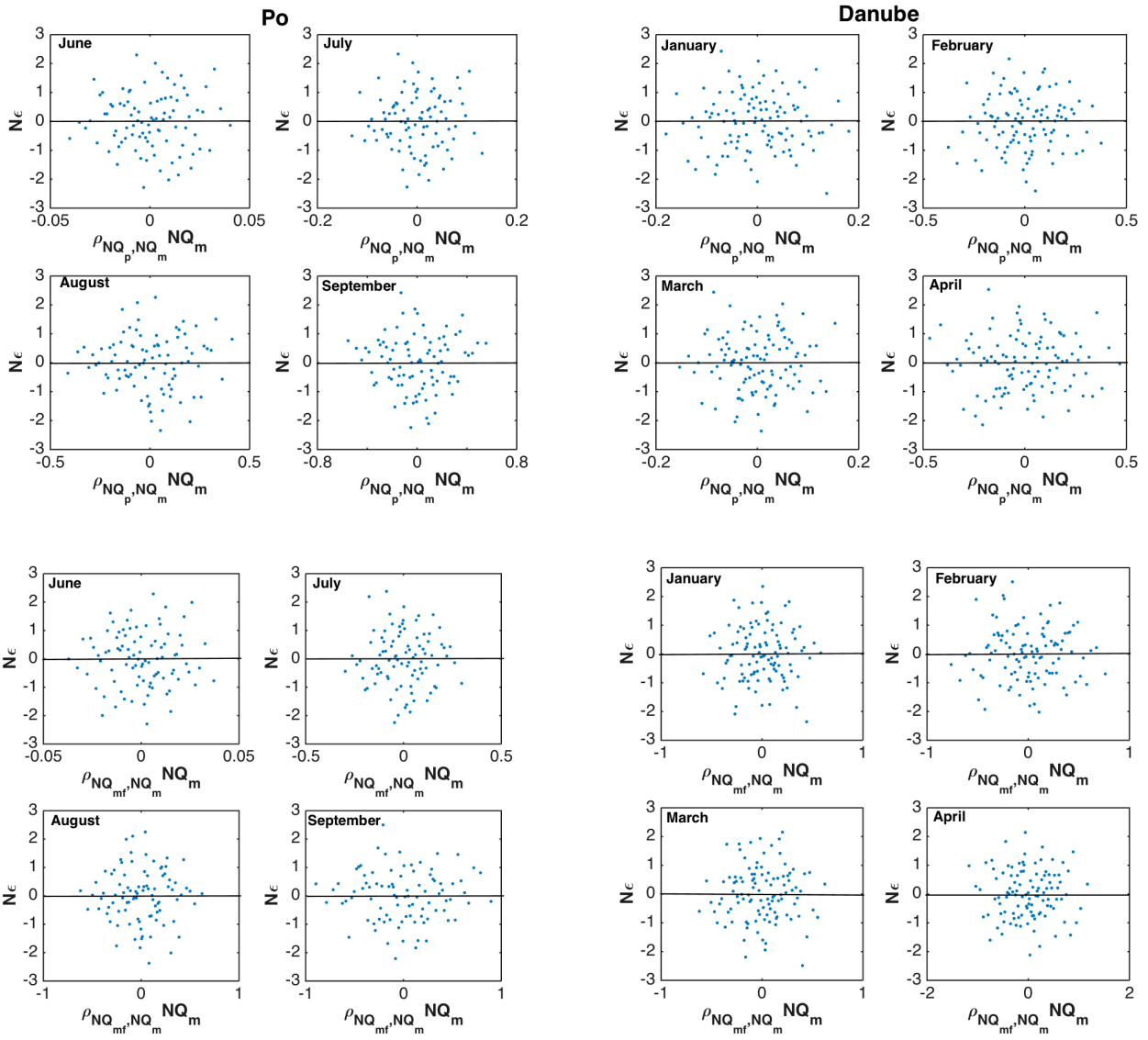
22

Figure 4. Residual plot of the linear regression of NQ$_m$ on NQ$_p$ and NQ$_{mf}$ in the Po river (left) and upper Danube (right)
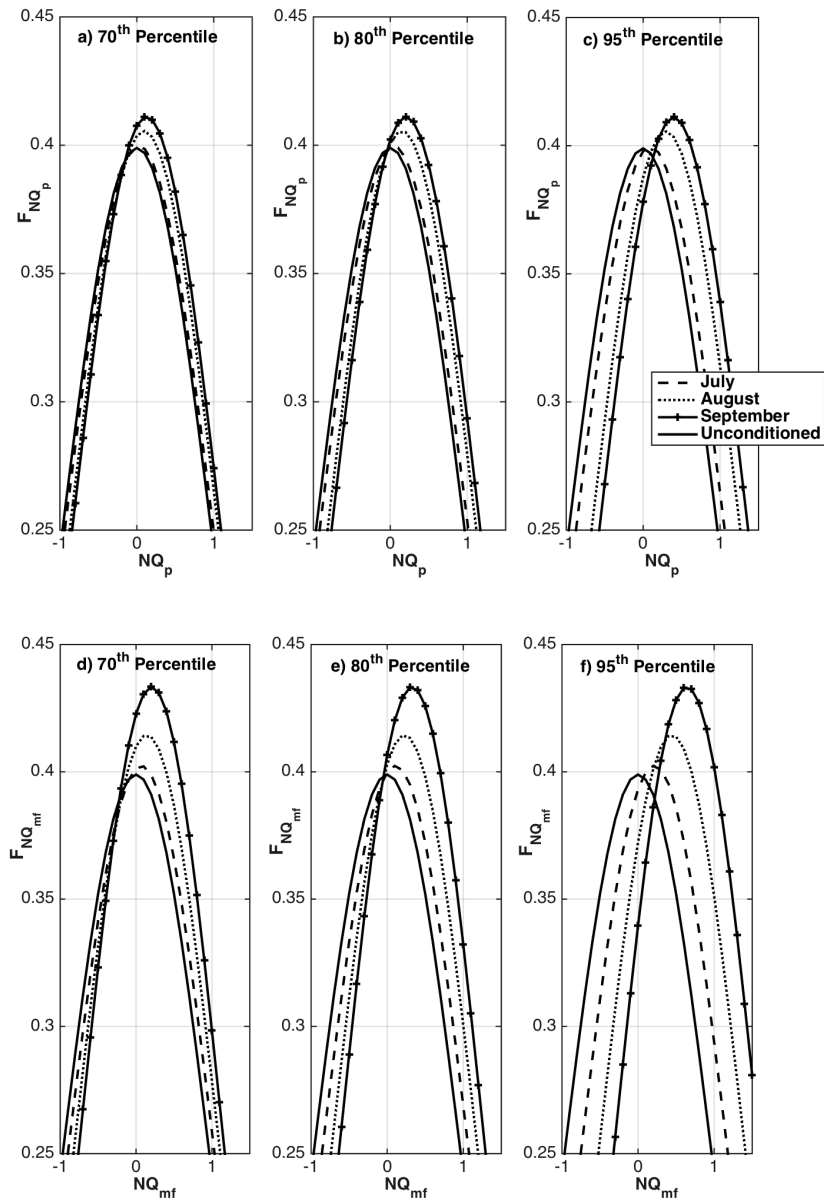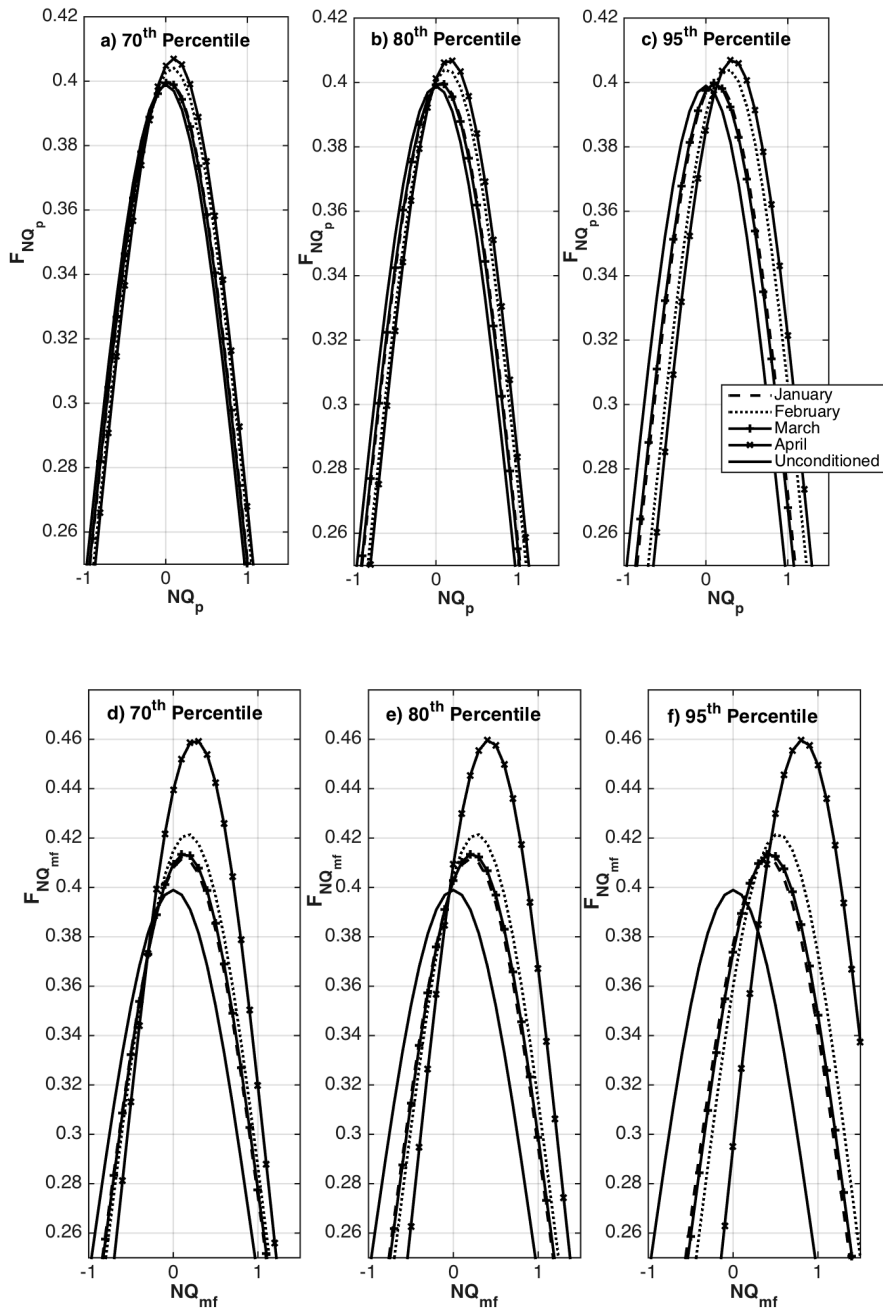
5

Figure 5. Probability distribution functions of the normalized dependent variables (NQ_p and NQ_{mf}) conditioned to the occurrence of the 70th, 80th and 95th percentiles of the normalized variables in the pre-flood season in the Po river.
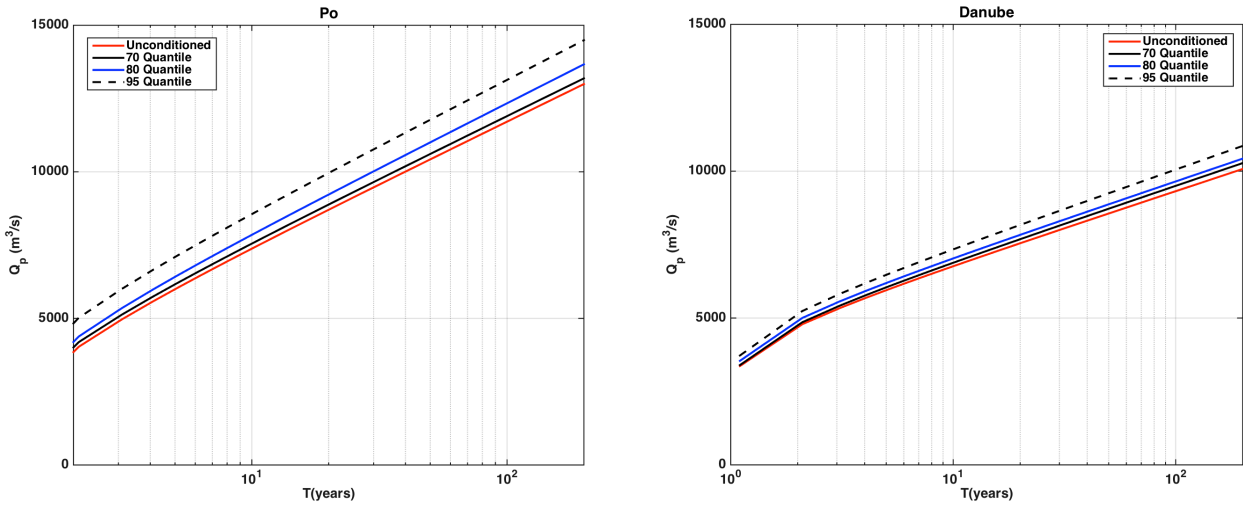
5

Figure 6. Probability distribution functions of the normalized dependent variables (NQ_p and NQ_{mf}) conditioned to the occurrence of the 70th, 80th and 95th percentiles of the normalized variables in the pre-flood season in the upper Danube.

Figure 7. Peak flows in the flood season (Oct-Nov in the Po, May-July in the upper Danube) *vs* return period modeled through the EV1 distribution function. Quantiles refer to mean flows higher than usual in the previous month
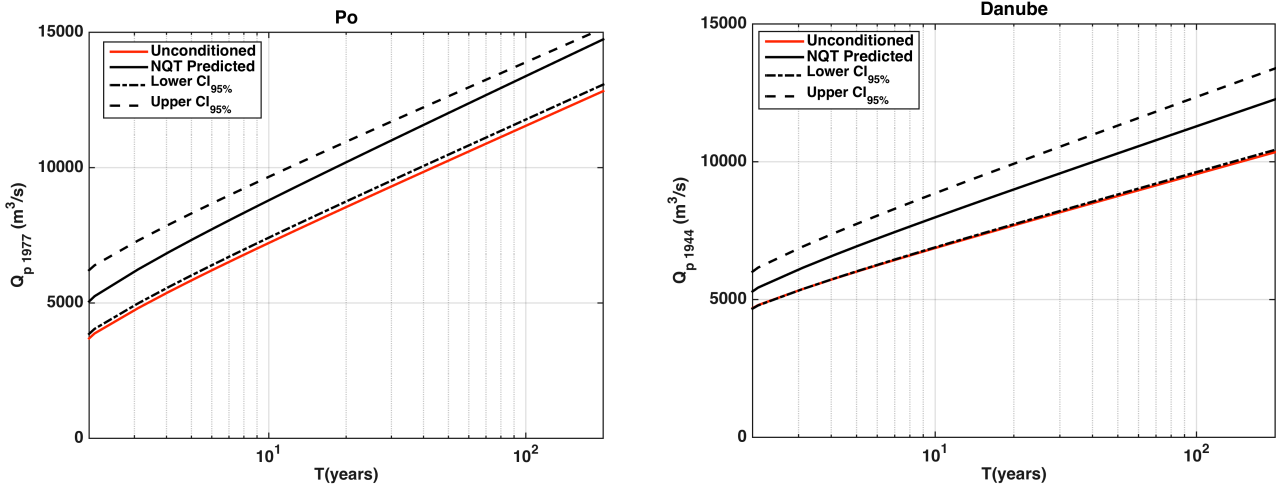

Figure 8. Leave-one-out cross validation. Unconditioned EV1 probability distribution of peak flows for the year with the wettest pre-flood season (1977 in the Po, 1944 in the upper Danube) along with conditioned distributions with related 95% confidence bands.

Table 1. Data description of observed time series. Descriptive statistics are given for non-deseasonalized data.

	Po	Danube
Observation period	1920-2009	1901-2007
Daily discharge		
Gauge location	Pontelagoscuro	Bratislava
Catchment area (km^2)	71 000	131 331
Mean (m^3 s^{-1})	1470	2053
Standard deviation (m^3 s^{-1})	1007	973
Fluvial regime	Pluvial regime. Two peak periods	Alpine regime. One peak in the summer
Monthly precipitation		
Number of weather stations	18	16
Mean (mm/month)	72	73
Standard deviation (mm/month)	17	37
Monthly temperature		
Number of weather stations	12	11
Mean (°C)	12.9	7.9
Standard deviation (°C)	7.5	7.2

5

Table 2. Estimated H values on deseasonalized data series applying R/S statistic (R/S), Aggregated Variance Method (AV), and Differenced Variance Method (DV)

	R/S	AV	DV
Po (Pontelagoscuro)			
Daily Q	0.81	0.74	0.94
Monthly Q	0.76	0.62	0.80
Monthly P	0.61	0.59	0.60
Monthly T	0.64	0.80	0.90
Danube (Bratislava)			
Daily Q	0.80	0.71	0.86
Monthly Q	0.75	0.54	0.79
Monthly P	0.56	0.36	0.56
Monthly T	0.61	0.76	0.70

Table 3. Pearson's cross-correlation coefficient and its 95% confidence interval between both, NQ_p and NQ_{mf} , and NQ_m for varying antecedent monthly flow. Flood season in Po: October-November. Flood season in Danube: May-July

Po			Danube		
Month	$\rho(NQ_m, NQ_p)$	$\rho(NQ_m, NQ_{mf})$	Month	$\rho(NQ_m, NQ_p)$	$\rho(NQ_m, NQ_{mf})$
September	0.24 (0.04, 0.43)	0.39 (0.20, 0.55)	April	0.20 (0.01, 0.38)	0.50 (0.34, 0.63)
August	0.18 (0, 0.37)	0.27 (0.07, 0.45)	March	0.06 (0, 0.25)	0.26 (0.07, 0.43)
July	0.06 (0, 0.26)	0.13 (0, 0.33)	February	0.16 (0, 0.34)	0.32 (0.14, 0.48)
June	0.02 (0, 0.23)	-0.02 (-0.23, 0)	January	0.07 (0, 0.26)	0.25 (0.06, 0.42)
May	-0.06 (-0.26, 0)	-0.05 (-0.25, 0)	December	-0.002 (-0.19, 0)	0.17 (0, 0.35)
April	-0.13 (-0.33, 0)	-0.07 (-0.27, 0)	November	0.05 (0, 0.24)	0.09 (0, 0.27)
March	-0.18 (-0.37, 0)	-0.12 (-0.32, 0)	October	0.13 (0, 0.31)	0.10 (0, 0.28)
February	-0.04 (-0.25, 0)	-0.05 (-0.25, 0)	September	-0.07 (-0.26, 0)	-0.08(-0.27, 0)
January	-0.07 (-0.27, 0)	-0.07 (-0.27, 0)	August	-0.21 (-0.38, -0.02)	0.09 (0, 0.27)