Clustering Flood Events from Water Quality Time-Series using Latent Dirichlet Allocation Model

Alice Aubert, Romain Tavenard, Rémi Emonet, Alban de Lavenne, Simon Malinowski, Thomas Guyet, René Quiniou, Jean-Marc Odobez, Philippe Mérot, Chantal Gascuel

Clustering flood events from water quality time-series using Latent Dirichlet Allocation model

Aubert, A.H.[1], Tavenard, R.[2], Emonet, R.[2], de Lavenne, A.[1], Malinowski, S.[3], Guyet, T.[3], Quiniou, R.[3], Odobez, J.-M.[2], Merot, Ph.[1], Gascuel-Odoux, C.[1]

[1]INRA, UMR 1069, Soil Agro and HydroSystem, Rennes, France
[2]IDIAP Research Institute, Martigny, Switzerland
[3]Agrocampus Ouest, UMR 6074, IRISA, Rennes, France
[4]INRIA, UMR 6074, IRISA, Rennes, France

Abstract

To improve hydro-chemical modeling and forecasting, there is a need to better understand flood-induced variability in water chemistry and the processes controlling it in watersheds. In the literature, assumptions are often made, for instance, that stream chemistry reacts differently to rainfall events depending on the season; however, methods to verify such assumptions are not well developed. Often, few floods are studied at a time and chemicals are used as tracers. Grouping similar events from large multivariate datasets using principal component analysis and clustering methods helps to explain hydrological processes; however, these methods currently have some limits (definition of flood descriptors, linear assumption, for instance). Most clustering methods have been used in the context of regionalization, focusing more on mapping results than on understanding processes. In this study, we extracted flood patterns using the probabilistic Latent Dirichlet Allocation (LDA) model, its first use in hydrology, to our knowledge. The LDA method allows multivariate temporal datasets to be considered without having to define explanatory factors beforehand or select representative floods. We analyzed a multivariate dataset from a long-term observatory (Kervidy-Naizin, western France) containing data for four solutes monitored daily for 12 years: nitrate, chloride, dissolved organic carbon, and sulfate. The LDA method extracted four different patterns that were distributed by season. Each pattern can be explained by seasonal hydrological processes. Hydro-meteorological parameters help explain the processes leading to these patterns, which increases understanding of flood-induced variability in water quality. Thus, the LDA method appears useful for analyzing long-term datasets.

Keywords: Water quality; time series analysis; flood events; probabilistic model; temporal pattern extraction

Keypoints:

- Flood water-chemistry studied with the LDA method
- LDA is a promising new pattern-extraction method for hydrologists
- Time distribution of water-chemistry flood patterns is explained by hydrological processes

1 Introduction

Studying floods has been a major issue in hydrological research for years. Given climate change implications, it will probably remain a hot topic for scientists [15]. Floods are often considered as an unusual amount of water that places a vulnerable environment at risk, but they also induce changes in water quality (*e.g.*, suspended sediments, sediment-bound chemicals). A pulse of rainfall transmits a signal in rivers that is visible in both water level and chemistry.

Numerous studies have focused on linking water quality and hydrological processes, both at base-flow and during floods (*e.g.* [6, 7, 8, 14, 23, 24]), but only a few have studied several solutes at a time (*e.g.* [5, 25, 30]) and analyze long-term time series. The present study follows previous multi-solute research in which base-flow processes are well-described [2, 3], but flood processes are only hypothesized. There is a need to better understand flood-induced variability in water chemistry; therefore, developing new methods is a current research topic.

More specifically, we aimed to test the assumption that stream chemistry reacts differently to rainfall events according to the season [2, 17, 22, 24] and the weather of the hydrological year [3, 23]. This assumption is made frequently, since few methods have been developed to identify chemical patterns and analyze when in the hydrological cycle they occur. Until now, researchers studying flood-induced variability in water chemistry have used a variety of methods, adapting them to the available data (*e.g.*, number of flood events, solutes monitored, sampling rate).

In some flood studies, water composition is seen as a hydrological marker that can reveal water transfers in a watershed [9, 19]. Discharge data are usually measured continuously, while chemistry data are usually collected automatically given a time lag or a discharge variation rate. Three methods have been developed. First, the hysteresis method, a common way to study water chemistry dynamics during floods [37, 19], consists of a discharge-concentration plot that can be used to identify floods pattern. The hystereses are translated into transport processes, i.e. chemical availability and spatial origins. Second, temporal graphs are used (*e.g.*, discharge and concentrations over time). Qualitative statistics such as mode, spread and skewness result from these studies. These two methods are often used when a single element is known to be exported

during flood events (*e.g.*, dissolved organic carbon (DOC), but more often suspended sediments or a chemical element bound to them, such as phosphorus). Third, end-member mixing analysis is another valuable tool for studying water pathways in a watershed [9]. By selecting certain water compartments and tracers and then monitoring tracer dynamics during a few storms, the contribution of the water compartments to stream water composition during storms can be estimated. To date, only a small number of flood events have been studied at once (in 1996, Durand and Torres focused on two events, in 2007, Lefranois *et al.* studied 142 floods from two catchments). To draw more general conclusions, it would be interesting to study more floods at the same time. The development of automated probes, generating large water chemistry datasets, will enable that and increase the need for automated methods to deal with large numbers of floods easily.

Other flood studies aim to group similar events, which are then explained in terms of hydrological processes. To this end, hydrologists analyze large multivariate datasets using principal component analysis (PCA), canonical discriminant analysis (CDA) and various clustering methods (CMs). Initially, these methods are used to reduce the volume of data (either by shrinking dimensionality or by representing observations with a smaller representative set) to make datasets more interpretable [29]. In papers focusing on flood-induced variability in water chemistry, PCA or CDA are most often used after defining a set of variables describing phenomena thought to be important. For instance, to use CDA, [32] identified four variables defining the weather (*e.g.*, total rainfall, average rainfall intensity), four hydrological variables describing the floods (*e.g.*, average discharge), three variables describing sediment transport (*e.g.*, maximum concentration during the flood, mean concentration) and seven explanatory factors describing conditions prior to floods (*e.g.*, rainfall and discharge prior to floods, interpolated soil moisture) as factors potentially influencing suspended-sediment dynamics during floods. [36] identified 24 descriptive variables for a PCA. In these studies, these variables describe flood events statically causing the loss of time dimension: instead of time series, a set of descriptive variable is used. PCA and CDA group variables by linearly combining them and then use the principal components to explain the cloud of observations. Correlations between factors are used to explain processes and lead to physically based interpretations [26]. Observations governed by the same factors can themselves constitute a group. PCA, based on many data, enables floods to be studied without considering solutes as a hydrological tracer. However, defining descriptive variables that seem to influence floods before analyzing the data may bias the results. Note that in completely different contexts, PCA can be run on temporal datasets, as [1] did when looking for spatial long-term water-quality classification. Another limit of PCA is its assumption of linearity. In quantitative hydrology, input variable selection algorithms are also used to select the most relevant input variables with respect to one, or more, output variables over pass the limits of PCA [31, 11]. To better understand flood-process dynamics, grouping methods are needed that retain the time dimension within results and do not require users to select potentially influential variables beforehand.

CMs are one solution to overcome this limitation, but an initial literature search revealed little use of CMs in studies of floods and water quality. Therefore, we widened the review to studies of qualitative and quantitative hydrology and found that clustering methods are used to regionalize models [21, 35]. CMs create groups of observations that are based on temporal datasets with minimum within-group and maximum between-group differences and are diverse; examples include k-means cluster analysis, hierarchical clustering, artificial neural networks (such as self-organizing maps) and interval clustering [1, 12, 13, 16, 26, 34, 38] and are sometimes based on PCA axes [33]. In these studies, however, results do not show the dynamics of the processes.

Another solution is to group temporal datasets without using descriptive variables, providing results in which dynamics can be observed based on pattern recognition [27]. This method has been used to generate synthetic streamflows [28]. In the present study, we used the probabilistic Latent Dirichlet Allocation (LDA) model [4], which also groups temporal data into patterns without using descriptive variables and provides results that retain the dynamics of processes. To our knowledge, this is its first use in hydrology. The LDA method analyses multivariate datasets without the need to define explanatory factors beforehand or select representative floods. The common k-means algorithm, applied to the temporal dataset, was used to assess the performance of the LDA method. By clustering floods from an environmental observatory, we aimed to test the assumption that stream chemistry reacts differently to rainfall events according to the season [2, 17, 22, 24] and the weather of the hydrological year [3, 23].

After describing the long-term observatory of Kervidy-Naizin and the available multivariate dataset collected over 12 years on this study site, we present methods in detail for the LDA model and briefly for the k-means algorithm. Results of LDA runs are presented in the Results and discussion section. One run (four patterns based on solute concentration only) is described in detail and its results are compared with those from k-means clustering. Then, we focus the discussion on what the resulting flood patterns indicate about the link between water chemistry and hydrological processes. We finish by discussing the new application of the LDA method to hydrology.

2 Material and method

2.1 Study site and available data

The Kervidy-Naizin watershed, located in western France (Brittany: 48°N, 3°W), contains a critical-zone observatory (ORE-AgrHys). This headwater watershed of approximately 5 km^2 is drained by a second order stream that occasionally dries in summer (Fig. 1). Data, metadata and scientific articles about the watershed are available on the ORE-AgrHys website (`http://www7.inra.fr/ore_agrhys_eng`).

Watershed topography is fairly flat (maximum slope = 5%) and its eleva-

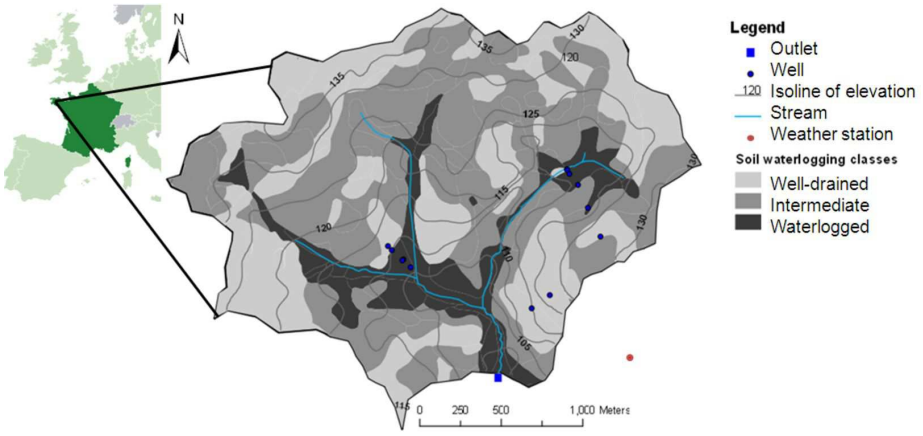

Figure 1: Map of the study site (Kervidy-Naizin, France).

tion ranges from 98-140 m above sea level. Kervidy-Naizin is an agricultural watershed with intensive animal farming. In 2010, 20% of its surface area was covered by cereals, 30% by maize and 20% by temporary or permanent pastures. The climate is temperate with oceanic influence, with a mean maximum daily temperature of 15.1°C (2000-2012). Mean annual rainfall is 818.5 mm, with the maximum and minimum monthly mean reached in November (around 100 mm) and in June (around 40 mm), respectively. The annual stream-specific discharge is approximately 350 mm year -1. Shallow groundwater, the main contributor to discharge, develops in an unconsolidated layer of weathered material up to 30 m thick, covering Upper Proterozoic schist. Soils are silty loams, and those on hillslopes are well-drained (Dystric Cambisols and Luvisols). Conversely, soils in the lowest zones are often saturated by groundwater rising to the surface (Epistagnic Luvisols and Epistagnic Albeluvisols) and constitute wetlands.

This study analyses data from January 2000 to August 2012 (slightly more than 12 hydrological years with highly variable hydrology and weather (Table 1)). Discharge was monitored once per minute at the outlet with a gauging station including a float-operated sensor and a data logger (Thalimdes OTT). The weather station at Kervidy (Cimel Enerco 516i) is located approximately one km from the outlet. It records, among other things, hourly rainfall and air temperatures. Groundwater level (measured every 15 minutes) from one hillslope well monitored with pressure probes was also used. Raw discharge data were used to detect flood dates. To keep the same time step as water chemistry data, daily mean discharge, groundwater depth and cumulative rainfall were used to explain patterns.

Stream water was manually sampled daily at approximately the same hour (5PM local time), without specific sampling during floods. These instantaneous grab samples were immediately filtered (0.2 m) and stored in the dark at 4°C in propylene bottles filled to the top. Samples were analyzed at most

	Hydrological index	Yearly rainfall (mm)	Yearly mean temperature (°C)
2000-01	2.4	1323.0	15.7
2001-02	0.7	662.5	14.1
2002-03	1.4	784.5	15.2
2003-04	0.9	864.0	16.0
2004-05	0.3	471.5	15.9
2005-06	0.5	609.5	15.9
2006-07	1.4	916.0	17.1
2007-08	1.0	873.5	13.7
2008-09	0.9	798.5	12.9
2009-10	1.3	877.0	14.6
2010-11	0.7	829.5	14.1
2011-12	0.5	812.0	16.0

Table 1: Hydro-meteorological conditions of the 12 years of study, showing high variability.

two weeks after collection. During one hydrological year only (2002-2003), sampling frequency was decreased to once every 2-4 days. Nitrate, chloride and sulfate concentrations were measured by ionic chromatography (DIONEX DX 100). DOC concentrations were measured with a total organic carbon analyzer (Shimadzu TOC 5050A).

2.2 Flood detection method

In keeping with the choice of fully-automated data treatment, flood events were detected automatically and defined as a rapid increase in discharge (Table 2). Each data point was considered along with the following six data points. When these seven measures agreed with the definition of a flood (increase over seven points with rate thresholds), the next measure was considered and the same test is applied. This continued until the definition was no longer satisfied. The potential flood event was then tested to determine whether its increases in speed and volume were sufficiently large. If so, the event was registered [18]. The threshold parameters were adjusted manually, using graphical validation.

We then extracted data from the solute concentration time-series for a 12-day period that included the two days before the day that each flood event began (identified during the automated treatment described previously) and the nine days after. Therefore, the present work considered floods within their hydrological context, including initial conditions (pre-event context), the flood signal, and the return to a stable state (post-event context) whose length varies greatly. Hereafter, these 12-day long periods are referred to simply as "floods".

Name	Threshold	Unit	Definition
PARAM_Speed1	0.0075/60	%Q.min^{-1}	Rate / speed (relative increase in discharge between two consecutive measurements) to initiate flood rising limb
PARAM_Speed2	0.001/60	%Q.min^{-1}	Rate / speed during flood rising limb
PARAM_NbLook	6	-	Number of measurements considered at each iteration
PARAM_deltaTdim1	72×60	min	Minimum time necessary to start a new rising limb
PARAM_deltaEvent	1.2	%Q	Minimum increase in discharge during the rising limb of the potential flood event (peak Q value / first Q value)
PARAM_Volume	0.005	mm	Minimum flow volume during the rising limb of the flood event, approximated by the area under the rising limb
PARAM_Intensity	0.004	mm.min^{-1}	Flow volume divided by the duration of the rising limb.

Table 2: Parameters used to detect floods in the Kervidy-Naizin dataset.

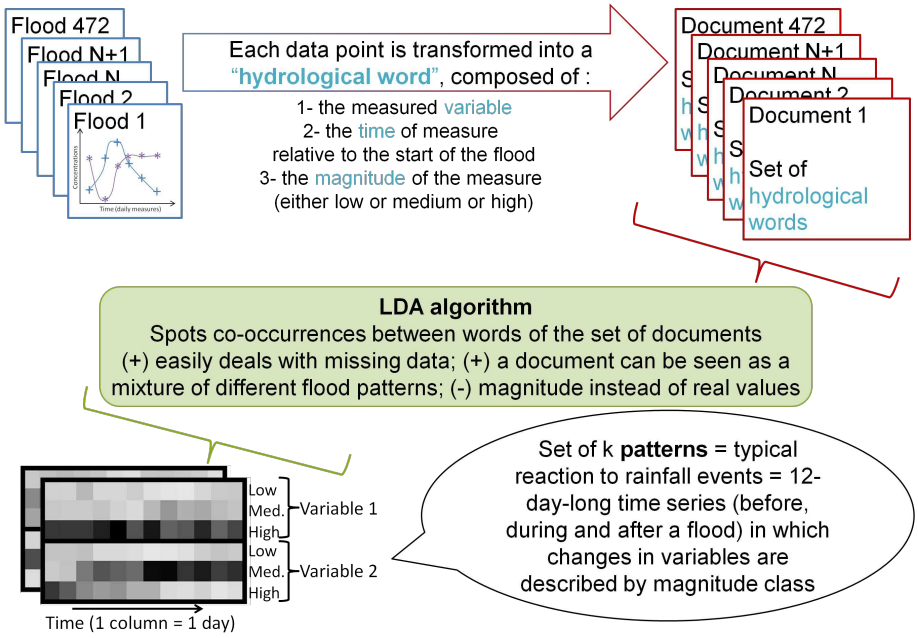

Figure 2: Diagram of the Latent Dirichlet Allocation (LDA) method.

2.3 LDA and clustering methods

2.3.1 LDA method

We used the generative model LDA [4] to mine recurrent sequential patterns from the flood dataset (Fig. 2). Generative models are probabilistic models that mimic the way data arise in observed data (documents). LDA is one such model and explains observations by hidden groups (latent topics). LDA was first introduced in the text-mining community, in which observations are words that occur in documents and depend on a small number of topics attached to the document. The aim of the method is to exhibit latent topics, which are distributions of word frequency, among all documents.

Given a collection of documents, LDA performs the following generative process (Fig. 3; variables in Table 3):

- For each topic k ($1 \leq k \leq K$):
  - Draw a distribution ϕ_k of words in topic k from a Dirichlet distribution of parameter β
- For each document i ($1 \leq i \leq N_d$):
  - Draw a distribution $heta_i$ of topics in document i from a Dirichlet distribution of parameter α

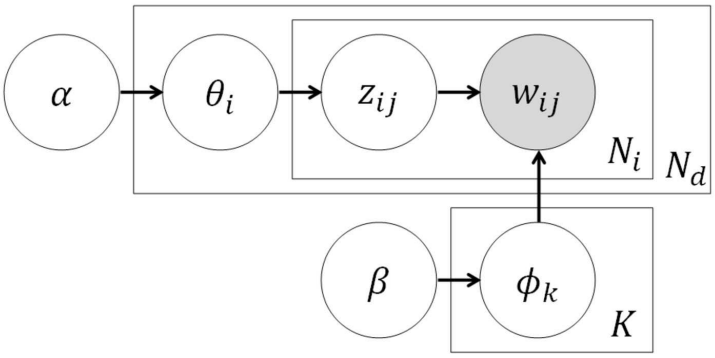

Figure 3: Diagram of the Latent Dirichlet Allocation generative process (variables are given in Table 3). α: vector of prior weights of topics in documents; *heta*: matrix describing the distribution of topics in documents; z: topic associated to an observation in a document; β: vector of prior weights of words in topics; ϕ: matrix describing the distribution of words in topics; w: word associated to an observation in a document.

- For each observation j ($1 \leq j \leq N_i$) in document i:
  - Draw a topic z_{ij} from a K-dimensional categorical distribution with probabilities $heta_i$
  - Draw a word w_{ij} from a W-dimensional categorical distribution with probabilities $\phi_{z_{ij}}$

Among these variables, only words w_{ij} are observed. Thus, parameters *heta*, ϕ and z have to be learned during inference. For the sake of simplicity, we chose to use the well-known variational inference algorithm, as in the original application of LDA to text mining [4]. For detailed information on this point, see [10].

To study floods with this model, we first defined what data to use as documents and words. From each event, a "flood document" was generated (Fig. 2). It contained a set of "hydrological words", each of which was composed of (i) the variable under consideration, (ii) the day on which the variable was measured (relative to the beginning of the flood), and (iii) the quantized measurement (i.e. low, medium or high magnitude). Unlike [4], who ignored word order, we chose to include the time dimension in words to identify temporal dynamics in floods. To distinguish our topics from those in other studies in which no time information is embedded, we refer to them as "patterns".

Flood documents (i.e. lists of hydrological words containing quantized variables sampled around the flood date) were used as input data to the LDA algorithm along with the required number of patterns to be extracted (K) —(see Appendix for a simplified example). LDA returned a set of K patterns that represented the most common flood behaviors mined from the flood documents.

Name	Type	Definition
K	Integer	Number of topics
W	Integer	Number of words in the dictionary
N_d	Integer	Number of documents
N_i	Integer	Number of observations in the i-th document
ϕ	$K \times W$ matrix of probabilities (each row sums to 1)	Matrix whose K-th row corresponds to distribution of words in topic K
θ	$N_d \times K$ matrix of probabilities (each row sums to 1)	Matrix whose i-th row corresponds to distribution of topics in document d_i
$\alpha = \{\alpha_1, \ldots, \alpha_K\}$	Vector of positive real values	Vector of prior weights of topics in documents (all α_k are set to 0.1 in our experiments). The lower these values, the sparser θ.
$\beta = \{\beta_1, \ldots, \beta_W\}$	Vector of positive real values	Vector of prior weights of words in topics (all β_w are set to 0.01 in our experiments). The lower these values, the sparser ϕ
z_{ij}	Integer between 1 and K	Topic associated with the j-th observation in document i
w_{ij}	Integer between 1 and W	Word associated with the j-th observation in document i

Table 3: Variables used to describe the generative Latent Dirichlet Allocation algorithm.

Class	Low	Medium	High
Rainfall (mm)	$<$ 1e-6	1e-6 – 5	$>$ 5
Specific discharge (mm d^{-}1)	$<$ 1e-6	1e-6 – 3.58	$>$ 3.58
Hillslope water table depth (m)	$<$ 1.86	1.86 – 3.86	$>$ 3.86
Mean temperature (°C)	$<$ 11	11 – 19	$>$ 19
Chloride	$<$ 25	25 – 35	$>$ 35
Sulfate	$<$ 7	7 – 11	$>$ 11
Dissolved organic carbon	$<$ 2.8	2.8 – 8.3	$>$ 8.3
Nitrate	$<$ 45	45 – 69	$>$ 69

Table 4: Boundaries of magnitude classes of hydrological parameters and solute concentrations (mg.l^{-}1) used in Latent Dirichlet Allocation runs.

Patterns are thus distributions of hydrological words by class. We represented them as 2D images in which the shade of gray represents the probability of the word occurring in the topic, denoted as $\phi_{(k,w)}$ (darker = higher). The x-axis in these images corresponds to the day within the 12-day period, while the y-axis represents variables and their magnitude classes. Quantization boundaries (Table 4) were set to encapsulate semantics in the observations when possible. For example, for the "Rainfall" variable, the "low" class was defined as no rainfall at all to separate out observations with a small amount of rain ("medium" class). When no measurement was available for a given variable on a given day (missing data), no word was generated. Thus, documents had different lengths, which posed no problem for LDA, which treats documents as word-frequency vectors (i.e. all N_i do not have to be equal). Note also that, running on quantized version of the data, LDA makes no assumption about linearity of the processes to be observed.

We first performed LDA requesting two patterns ($K = 2$) and using both solute-concentration and hydro-meteorological variables; however, model parameters specialized on the latter, giving little information about the former. Thus, we repeated the LDA but considered only solute-concentration variables. It identified a pattern (pattern 1) that was dominant for eight months of the year, suggesting that this pattern was over-represented. This motivated our last LDA run, in which four patterns were requested ($K = 4$) considering only solute-concentration variables. Results of this run confirmed our previous concern, since the 8-month period was covered by 2 patterns (Patterns 1 and 3).

2.3.2 Clustering Method

For comparison, we also analyzed the flood time-series using a well-known clustering algorithm: the k-means algorithm [20]. This algorithm aims to find groups (clusters) of similar elements from a dataset. In our case, the elements of the dataset were the 12-day solute time-series, and similarity between elements was calculated as the Euclidean distance. To deal with missing data, we used a modified version of the Euclidean distance that compared only the data

that were present in both time-series and averaged their distances. Input for the k-means algorithm is the number of clusters desired. The k-means algorithm iteratively assigns elements to clusters until it converges to a stable solution. The elements in a cluster are averaged to extract the cluster's "centroid". One difference between k-means and LDA is that the former uses real values of the series, whereas LDA uses magnitude classes.

We applied the k-means algorithm by requesting four clusters. Discarding floods with too many missing data left 364 for analysis. The four cluster mean behaviors were presented in a solute graphs (Fig. 7). To consider one cluster, one needs to check the lines of the same colors for the four solutes, because with CM, the real values of the parameters are used: they are not processed into classes. As solute concentrations were of different order of magnitude, we could not plot them in a single graph. Since the output of the k-means algorithm depends strongly on initial settings, 100 k-means algorithms with random initializations were run on the dataset, and we selected the clustering result that minimized the sum of squared (SSQ) distances. The SSQ is a classical index to evaluate the quality of clustering. Low SSQ values ensure that the clusters formed are compact around their centroids.

3 Results and discussion

4 The data set

Automatic flood-detection identified 472 12-day periods over the 152 months of the time-series. As a reminder, "floods" in this study are 12-day periods in which the flood event occurs the third day. When a flood is receding, another flood event can occur, causing overlaps. Flood frequency (Fig. 4) was seasonal, with ≥ 50 floods per month from November to March (maximum = 85 in January) and ≤ 20 floods per month from June to September (minimum = three in September). Only 38 months of the time-series had no floods, when the stream dried up: March (three times), April and May (once each), June and July (five times each), August (eight), September (ten), October (five) and November (once). Drying of the stream was also the main reason for missing data.

4.1 Patterns obtained when applying LDA to hydrological time-series to cluster floods

4.1.1 LDA application

Two patterns based on solute concentrations and hydro-meteorological conditions The rainfall causing the flood events is easily visible on days 3 and 4. Temperature and mid-slope groundwater level strongly influenced differences in patterns, being medium for pattern 0 and low for pattern 1 (Fig. 5a). Given the marked seasonality of these two variables, LDA may have focused on them, giving little information about other variables.

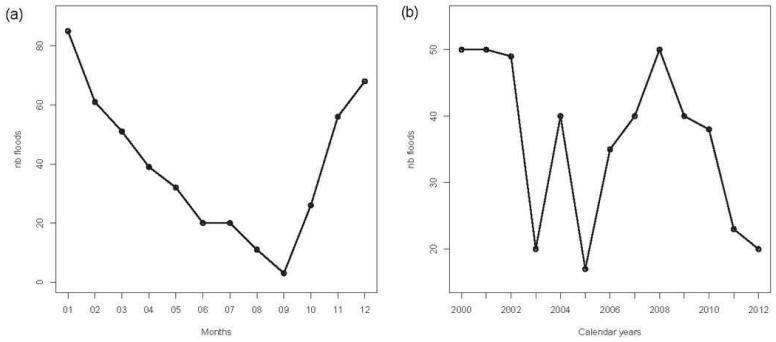

Figure 4: Flood frequency (n = 472) (a) monthly, (b) yearly.

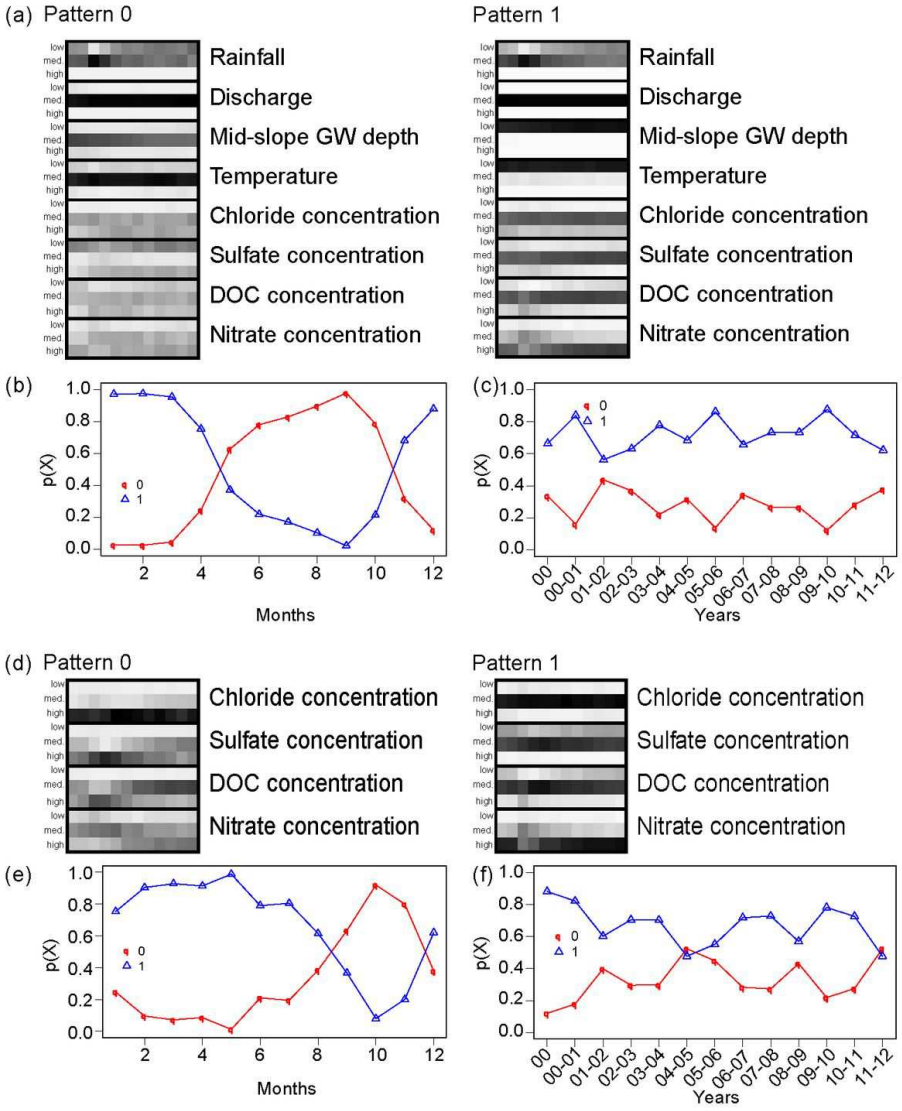

Figure 5: Patterns obtained with K = 2 and solute-concentration parameters (a) with or (d) without hydro-meteorological parameters and (b and e) their associated monthly and (c and f) yearly distributions. The x-axis in the pattern images corresponds to the day within the 12-day period, while the y-axis represents variables and their magnitude classes. DOC: dissolved organic carbon; GW: groundwater.

Two patterns based on solute concentrations only The patterns obtained clearly differed for all solutes (Fig. 5d). Pattern 0 showed slight dilution of high chloride concentration and marked increase in DOC and sulfate concentrations lasting about two days following the event, and slight dilution of nitrate concentration in a context of gradually increasing concentrations. In theory, such patterns correspond to reactions observed at the start of the hydrological year. The occurrence of pattern 0 mainly in October and November confirmed this (Fig. 5e). Pattern 1 showed slight dilution of medium chloride concentration, slight increase in sulfate concentration, a peak in DOC concentration on the day of the event, and clear dilution of high nitrate concentration. This pattern, occurring during the eight months of the year when flow was high (winter and spring), was complementary to pattern 0. The driest of the 12 years (2004-2005) behaved differently (Fig. 5f). Having only one group representative of eight months of the year seemed poor, so we requested four groups.

Four patterns based on solute concentrations only When we requested four groups, the LDA method satisfactorily extracted patterns from the multivariate chemical signature of floods within their contexts (Fig. 6). Pattern 0 showed slight dilution of high chloride concentration, marked increase in DOC and medium sulfate concentrations lasting about two days following the event, and slight dilution of nitrate concentration in a context of gradually increasing concentrations. Defining a 12-day window around floods enabled us to study short-term variation during the core of flood events, as well as differences between initial and final conditions over the 12 days. These conditions defined the contexts of floods, which in this pattern were stable concentrations of chloride, sulfate, and DOC, but increasing concentration of nitrate. In theory [2], such pattern corresponds to reactions observed at the start of the hydrological year. The occurrence of pattern 0 mainly in October and November confirmed this (Fig. 6b).

Pattern 1 showed no clear change in medium chloride concentration, increase in low sulfate concentration, a peak in DOC concentration on the day of the event, taking several days to return to low concentration, and clear dilution of high nitrate concentration. In theory, pattern 1 corresponds to high flow periods when the chloride stock was empty. The occurrence of pattern 1 predominantly from April to August confirmed this (Fig. 6b).

Pattern 2 showed dilution of high chloride concentration, slight increases in medium sulfate and DOC concentrations, and clear dilution of high nitrate concentration lasting several days. In theory, pattern 2 corresponds to high flow periods when chloride, DOC and sulfate stocks remain high. The occurrence of pattern 2 mostly from October to January confirmed this, even if it was not the predominant pattern (Fig. 6b).

Pattern 3 showed slight dilution of medium chloride concentration, almost no change in medium sulfate concentration, a peak in medium DOC concentration, and clear dilution of high nitrate concentration lasting several days, in a context of increasing concentration. In theory, pattern 3 corresponds to the start of

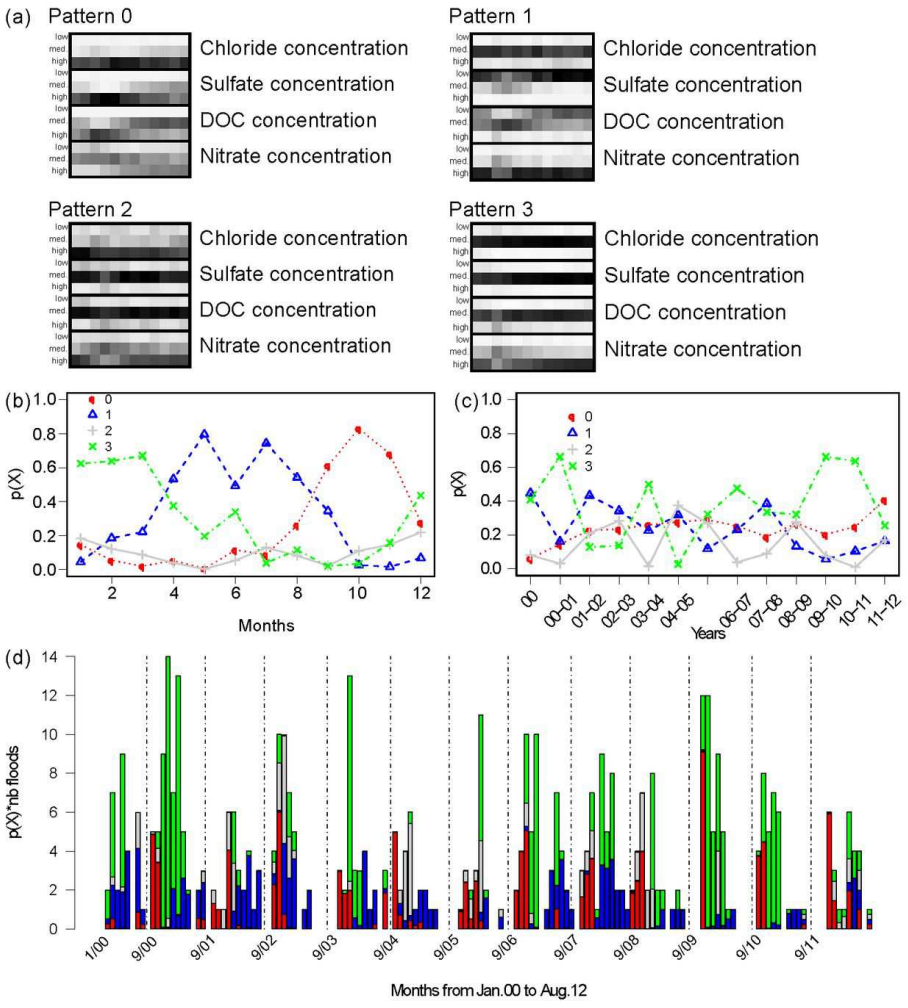

Figure 6: Patterns obtained (a) with k = 4 with solute-concentration parameters alone and their associated (b) monthly and (c) yearly distributions and (d) monthly distribution over the 12 years weighted by the number of floods. The x-axis in pattern images corresponds to the day within the 12-day period, while the y-axis represents variables and their magnitude classes.

high-flow periods. The occurrence of pattern 3 predominantly from January to March confirmed this (Fig. 6b).

4.1.2 Comparison with the clustering method

The flood clustering results, obtained using the k-means algorithm asking for four clusters, are given in Fig. 7. Cluster A showed a lasting concentration (bell curve) of high chloride concentration, a peak in high DOC concentration, a peak in decreasing sulfate concentration, and continuous increase in nitrate concentration (which did not react to the flood). In theory, cluster A corresponds to the reaction observed at the start of the hydrological year. Cluster B showed no change in high chloride concentration, a peak in medium DOC concentration, no change in medium sulfate concentration and slight dilution of medium increasing nitrate concentration. In theory, Cluster B corresponds to a period when chloride is high and nitrate increases, which appears, like Cluster A, at the start of the hydrological year. Cluster C showed clear dilution of high chloride concentration, a peak in DOC concentration, a relatively clear peak in sulfate concentration, and clear dilution of high nitrate concentration. In theory, cluster C corresponds to floods that occur during high flow periods when the watershed is already rewetted. Cluster D showed no change in high chloride concentration, a slight peak in relatively low DOC concentration, no change in low sulfate concentration, and slight dilution of high nitrate concentration. In theory, cluster D corresponds to the end of high-flow periods. The monthly occurrence of cluster D from August to December confirmed this (Fig. 7).

Both methods were consistent in the sense that one LDA pattern corresponded to one CM cluster: pattern 0/cluster A, pattern 1/cluster B (particularly for DOC and sulfate), pattern 2/cluster C and pattern 3/cluster D.

4.2 Hydrological interpretation

First, the monthly distribution of patterns showed a clear seasonality of floods. The hydro-meteorological statistics for each pattern detailed the conditions in which the patterns occurred (Fig. 8). A pattern was linked to a set of hydrological conditions that influenced various hydrological processes.

Pattern 0 showed conditions occurring at the start of a hydrological year, i.e. increase in water-level depth at mid-slope, increase in discharge (on average, no falling limb is observed at a daily rate) and rainfall causing floods were higher, on average. Monthly pattern distribution over the 12 years also showed that pattern 0 occurred at the beginning of the hydrological year and was stronger after complete drying of the stream (Fig. 6d). Therefore, the flood patterns were explained by the following points: (i) nitrate gradually increased while upland groundwater connected with the stream; (ii) concentration peaks for DOC and sulfate were brief but high: after regenerating the previous summer, stocks were exported by the first connection of wetland groundwater with the stream; and (iii) chloride concentration is high and more likely to be higher after the flood: summer evapotranspiration concentrated chloride, creating a stock.

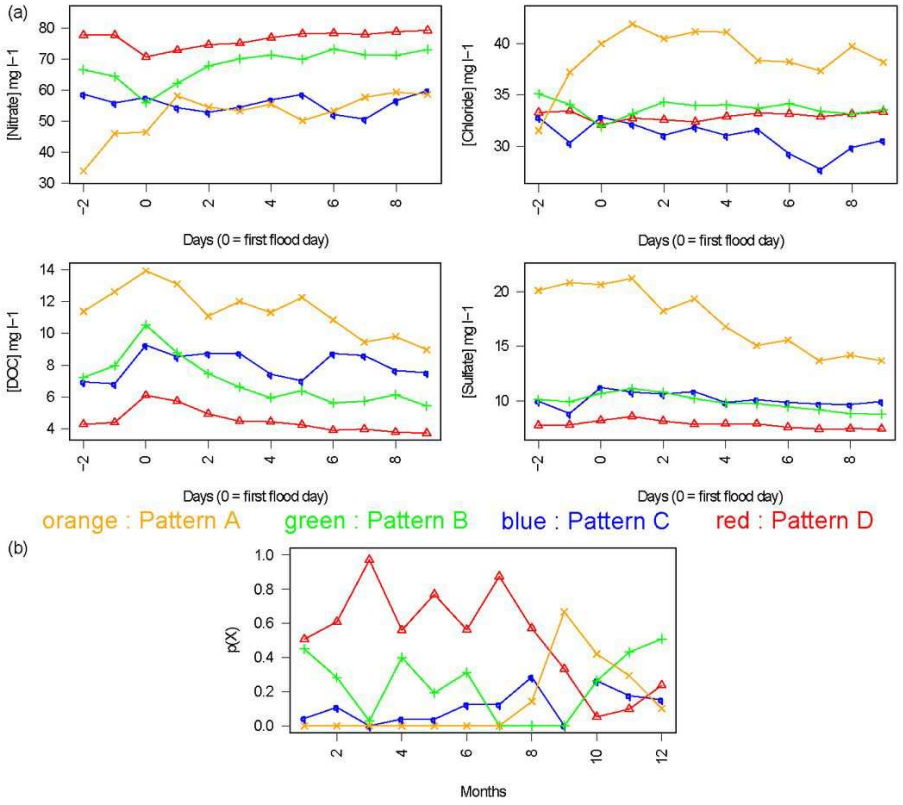

Figure 7: Clusters obtained (a) with the k-means clustering algorithm and (b) their associated monthly distributions.

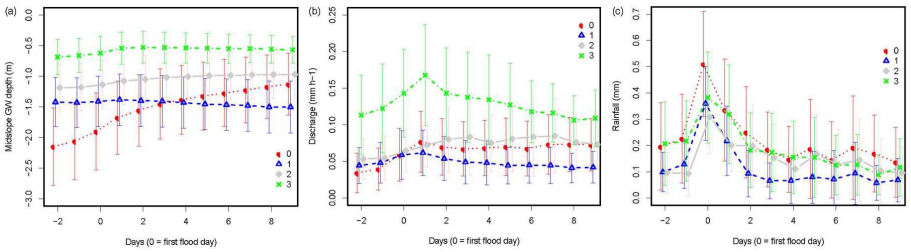

Figure 8: Statistics of the hydro-meteorological parameters corresponding to LDA pattern extraction with $K = 4$ and solute-concentration parameters alone: (a) mid-slope groundwater (GW) depth, (b) discharge, and (c) rainfall. Error bars equal 1 standard deviation.

Pattern 1 occurred when water level at mid-slope was approximately 1.5 m deep and discharge increase was smooth, in other words, after the high-flow period. Monthly pattern distribution over the 12 years also showed that pattern 1 occurred when the watershed was rewetted (after pattern 3) (Fig. 6d). Base concentrations of DOC and sulfate were low: stocks were depleted by previous floods but were still sufficient to provide solutes for export. Even though the upland groundwater table was lower, it remained connected and provided nitrate to the stream. The flood diluted high base concentrations: during floods, wetland groundwater, where nitrate was less concentrated, contributed more to the stream.

Pattern 3 was typical of high flow, showing the highest water-level depth and discharge. Monthly pattern distribution over the 12 years also showed that pattern 3 occurred in winters following a dry summer, in other words, following periods of high occurrence of pattern 0 (Fig. 5d). This could explain why chloride and sulfate dynamics were rather flat: most of the stock was already depleted by the first floods of the hydrological year.

Pattern 2 was less specific and showed the lowest probability of occurrence. Discharge and rainfall varied, and small rainfall amounts led to floods. Monthly pattern distribution over the 12 years (Fig. 5d) showed that pattern 2 looked like the counterpart of pattern 3: low when pattern 3 was high and vice-versa (Fig. 5c). Therefore, pattern 2 must occur in an already wet watershed and include overlapping floods in the nine days following the first flood. Pattern 2 could be called an opportunistic pattern lying between patterns 3 and 0. It was probably created because we requested four patterns. An additional LDA run asking for three patterns supported this hypothesis (not presented).

These analyses, focused on floods, support a previous study performed on the same watershed [2] and quantify several observations. For instance, the probability of chloride dilution occurring late in a high-flow period (pattern 3) appeared low. We highlight that such chloride dilution late in a high-flow period could occur in wet years but is less probable in years following a dry summer. Indeed, in wet years, summer chloride concentration in the wetland is not high and its stock is quickly exported, but in a dry summer, higher evapotranspiration increases chloride concentration, creating a large stock that requires more time to export when flow resumes. These latter years lead to increased sulfate and chloride export at the beginning of the hydrological year, since their stocks built up during the summer.

The yearly distribution of patterns (Fig. 6c), raised two questions: first, whether the increasing frequency of pattern 0 during the time-series indicated increasing contrast between summer and autumn, and second, whether pattern 2 predominated in 2004-2005 because it was the driest year. The first question can be answered by a longer time-series or comparison with other sites. For the second, yearly distribution of patterns for K = 2 (Fig. 5f) supports an affirmative response, confirming the assumed importance of annual meteorological conditions on flood patterns. As mentioned above, however, pattern 2 is an opportunistic pattern; it is possible that 2004-2005 had unique characteristics among the 12 years monitored.

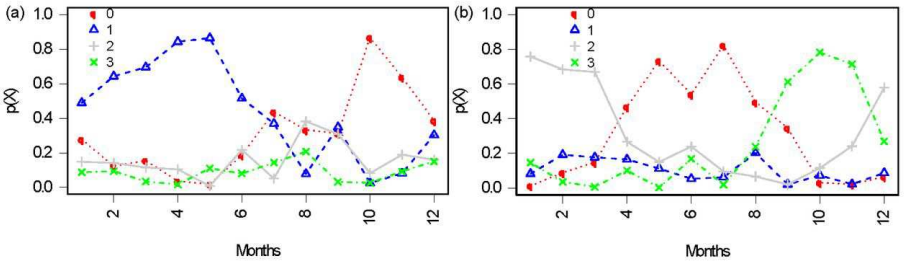

Figure 9: Monthly distribution of patterns obtained when applying LDA on groups of solutes known to be similar: (a) nitrate and chloride and (b) dissolved organic carbon (DOC) and sulfate.

We wondered how grouping the solutes in the manner suggested by [2] would influence results: nitrate and chloride, originating mostly from anthropogenic inputs creating excess within the watershed, vs. DOC and sulfate, originating from biogeochemical processes creating internal patches of production within the watershed. We ran LDA twice with two solutes. Monthly distribution of the patterns showed that nitrate and chloride were dominated by two patterns (one October-December, the other February-June) while DOC and sulfate were dominated by three patterns (one September-November, another December-March, and the last April-August) (Fig. 9). Monthly distribution of patterns obtained with only DOC and sulfate was similar to that obtained with all four solutes. DOC and sulfate appeared to have a stronger weight than nitrate and chloride when running LDA for the four solutes. This makes sense, since DOC and sulfate exports depended more on flood events, i.e. when a flood temporarily connected the stock in wetland groundwater with the stream.

Interesting points emerged to increase hydrological understanding. Knowing that the number of long-term watershed observatories is increasing and new probes develop quickly, longer time-series awaiting efficient analysis methods are becoming available. The LDA method could be one of them.

4.3 Methodological perspective

It is interesting to know when one or two LDA patterns explained nearly all (80-90%) floods of a document. We defined an "extra-pure" document as one for which only one pattern explained 90% of floods, and a "pure" document as one for which one pattern explained at least 50% and a second explained at least 30% of floods. Thus, an extra-pure document is, by definition, also considered pure. Most documents were pure, explained by two dominant patterns (Fig. 10a). Throughout the year, at least half of the documents were extra-pure, except in June, July and August. These three months had relatively few floods (20, 20 and 11, respectively) and were months of transition from high-flow to low- or no-flow periods (June, July and August were dry, respectively, 5, 5 and 7 times out of 12

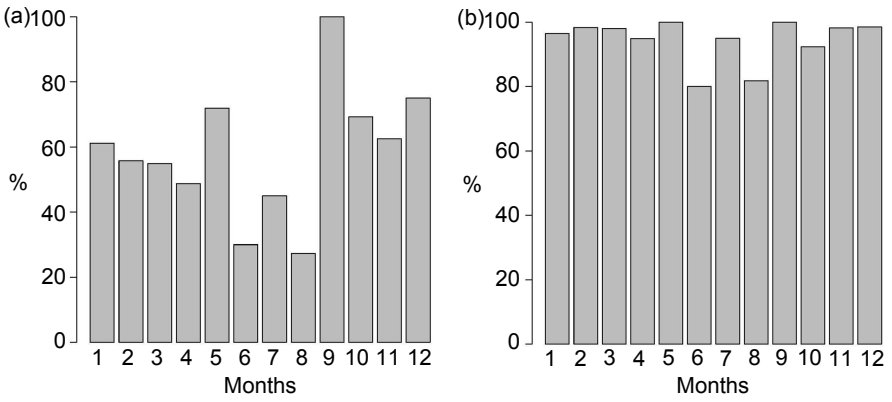

Figure 10: Histograms showing the percentage of (a) "extra-pure" and (b) "pure" documents per month. An extra-pure document is one explained $\geq$ 90% by one pattern. A pure document is one explained $\geq$ 50% by one pattern and $\geq$ 30% by a second pattern. Extra-pure documents are included in pure documents.

years). These reasons could explain why several patterns were needed to explain the floods of these months. However, September, also a transition month, was extra-pure, but this can be explained by specific conditions at the start of a hydrological year that induce a specific pattern. In addition, September was often dry (10 times) and had only three floods. LDA's ability to mix patterns within a document might explain the better monthly flood distribution obtained with LDA than with CM.

Another advantage of LDA versus the k-means algorithm, besides the ability to obtain mixed documents, is output stability. Indeed, we observed that results of LDA runs are more stable across runs than those of k-means.

Data quantization (i.e. the action of turning continuous data into categorical one) is a useful feature, even though precision decreases and the choice of thresholds has a large influence. Quantization decreases the importance of outliers and can add semantics to "hydrological words". Thus, the LDA method could be applied in studies of water quality by creating classes below or above the legal threshold. Spatial information can also be added to the word, making it possible to use LDA in regionalization studies.

In sum, (i) chemistry parameters alone were sufficient to obtain meaningful patterns with LDA; (ii) hydro-meteorological parameters were useful afterwards, however, to give hydrological meaning to patterns; (iii) defining limits to magnitude classes was important for obtaining clear patterns; (iv) LDA deals with missing data and overlapping events; (v) a pattern with poor ability to explain clusters can be easily identified and discarded; (vi) solutes whose export is strongly related to floods seemed more influential; (vii) by using mixed documents, LDA provides a finer time distribution than CM; and (viii) LDA

patterns were more easily readable than CM patterns.

In general, LDA enables accurate flood study and pattern extraction even if water chemistry is not specifically sampled during flood events, which is the case for most long-term hydrological observatories. Thus, it would be an interesting method for long-term datasets not focused on floods. The LDA method can be applied easily to data from large watersheds. Classes should be defined with special care according to the watershed, especially if the stream chemistry is dampened. Another strength of this method is that, unlike methods focusing on flood-induced water quality dynamics that have describe the event with statistics (time of rising limb, maximum and minimum values of variables thought to be of interest), the temporal dimension is maintained throughout the clustering, enabling dynamic processes to be studied.

5 Conclusion

We extracted flood patterns using the probabilistic LDA model, its first use in hydrology, to our knowledge. One major advantage of LDA is that it can cluster multivariate datasets without having to define explanatory factors beforehand or select representative floods. It would be particularly useful in watersheds where little is known about their hydrological functioning. LDA accurately extracted flood patterns that had a seasonal distribution, even though water chemistry was not specifically sampled during flood events. Objectively establishing links between water quality variations and hydrological processes improves understanding of variability induced by floods. Knowing that the number of long-term watershed observatories is increasing and new probes develop quickly, longer time-series awaiting efficient analysis methods are becoming available. The LDA method is one of them. It would also be a useful method for comparing and classifying watersheds based on their water quality and for evaluating whether water-quality meets regulations.

References

[1] Astel, A., S. Tsakouski, P. Barbieri, and V. Simeonov (2007), *Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets*, Water Research, 41(19), 4566-4578.

[2] Aubert, A.H., Gascuel-Odoux, C., Gruau, G., Akkal, N., Faucheux, M., Fauvel, Y., Grimaldi, C., Hamon, Y., Jaffrezic, A., Lecoz Boutnik, M., Molenat, J., Petitjean, P., Ruiz, L., Merot, Ph. (2013a), *Solute transport dynamics in small, shallow groundwater-dominated agricultural catchments: insights from a high-frequency, multisolute 10 yr-long monitoring study*, Hydrol Earth Syst Sci, 17(4): 1379-1391.

[3] Aubert, A.H., Gascuel-Odoux, C., Merot, P. (2013b), *Annual hysteresis of water quality: A method to analyse the effect of intra- and inter-annual climatic conditions*, J Hydrol, 478(2013): 29-39.

[4] Blei, D. M.; Ng, A. Y.; Jordan, M. I. (2003), *Latent Dirichlet allocation*, J Mach Learn Res, 3(4-5): 9931022.

[5] Burns, D. A., R. P. Hooper, J. J. McDonnell, J. E. Freer, C. Kendall, and K. Beven (1998), *Base cation concentrations in subsurface flow from a forested hillslope: The role of flushing frequency*, Water Resour Res, 34(12), 3535-3544.

[6] Chen, J., H. S. Wheater, and M. J. Lees (2002), *Identification of processes affecting stream chloride response in the Hafren catchment, mid-Wales*, J Hydrol, 264(1-4), 12-33.

[7] Creed, I. F., L. E. Band, N. W. Foster, I. K. Morrison, J. A. Nicolson, R. S. Semkin, and D. S. Jeffries (1996), *Regulation of Nitrate-N Release from Temperate Forests: A Test of the N Flushing Hypothesis*, Water Resour Res, 32(11), 3337-3354.

[8] Dawson, J. J. C., C. Soulsby, D. Tetzlaff, M. Hrachowitz, S. M. Dunn, and I. A. Malcolm (2008), *Influence of hydrology and seasonality on DOC exports from three contrasting upland catchments*, Biogeochemistry, 90(1), 93-113.

[9] Durand, P., Torres, J.L.J. (1996), *Solute transfer in agricultural catchments: The interest and limits of mixing models*, J Hydrol, 181(1-4): 1-22.

[10] Fox, C. W., and S. J. Roberts (2012), *A tutorial on variational Bayesian inference*, Artif Intell Rev, 38(2), 85-95.

[11] Galelli, S., and A. Castelletti (2013), *Tree-based iterative input variable selection for hydrological modeling*, Water Resour Res, 49(7), 4295-4310.

[12] Hannah, D.M., Smith, B.P.G., Gurnell, A.M., McGregor, G.R. (2000), *An approach to hydrograph classification*, Hydrol Process, 14(2): 317-338.

[13] Harris, N.M., Gurnell, A.M., Hannah, D.M., Petts, G.E. (2000), *Classification of river regimes: a context for hydroecology*, Hydrol Process, 14(16-17): 2831-2848.

[14] Hornberger, G. M., K. E. Bencala, and D. M. McKnight (1994), *Hydrological controls on dissolved organic carbon during snowmelt in the Snake River near Montezuma, Colorado*, Biogeochemistry, 25(3), 147-165.

[15] IPCC (2012), *Summary for Policymakers, in Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*, [Field,

C.B., V. Barros, T.F. Stocker, D. Qin, D.J. Dokken, K.L. Ebi, M.D. Mastrandrea, K.J. Mach, G.-K. Plattner, S.K. Allen, M. Tignor, and P.M. Midgley (eds.)]. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK, and New York, NY, USA, pp. 3-21.

[16] Kalteh, A. M., P. Hjorth, and R. Berndtsson (2008), *Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application*, Environmental Modelling & Software, 23(7), 835-845.

[17] Lambert, T., A.-C. Pierson-Wickmann, G. Gruau, J.-N. Thibault, and A. Jaffrezic (2011), *Carbon isotopes as tracers of dissolved organic carbon sources and water pathways in headwater catchments*, J Hydrol, 402(34), 228-238.

[18] de Lavenne, A., Cudennec, C. (2012), *Streamflow velocity estimation in GIUH-type approach: what can neighbouring basins tell us?*, Poster Presentation, EGU General Assembly, 22-27 April 2012, Vienna, Austria.

[19] Lefranois, J., Grimaldi, C., Gascuel-Odoux, C., Gilliet, N. (2007), *Suspended sediment and discharge relationships to identify bank degradation as a main sediment source on small agricultural catchments*, Hydrol Process, 21(21): 2923-2933.

[20] MacQueen, J. B (1967), *Some Methods for classification and Analysis of Multivariate Observations*, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press.

[21] Merz, R., and G. Bloschl (2003), *A process typology of regional floods*, Water Resources Research, 39(12).

[22] Molenat, J., and C. Gascuel-Odoux (2002), *Modelling flow and nitrate transport in groundwater for the prediction of water travel times and of consequences of land use evolution on water quality*, Hydrol Process, 16(2), 479-492.

[23] Molenat, J., C. Gascuel-Odoux, L. Ruiz, and G. Gruau (2008), *Role of water table dynamics on stream nitrate export and concentration. in agricultural headwater catchment (France)*, J Hydrol, 348(3-4), 363-378.

[24] Morel, B., P. Durand, A. Jaffrezic, G. Gruau, and J. Molenat (2009), *Sources of dissolved organic carbon during stormflow in a headwater agricultural catchment*, Hydrol Process, 23(20), 2888-2901.

[25] Mulholland, P. J., and W. R. Hill (1997), *Seasonal patterns in streamwater nutrient and dissolved organic carbon concentrations: Separating catchment flow path and in-stream effects*, Water Resour Res, 33(6), 1297-1306.

[26] Orwin, J.F., Smart, C.C. (2004), *Short-term spatial and temporal patterns of suspended sediment transfer in proglacial channels, small River Glacier, Canada*, Hydrol Process, 18(9): 1099-1085.

[27] Panu, U. S., T. E. Unny, and R. K. Ragade (1978), *Feature prediction model in synthetic hydrology based on concepts of pattern-recognition*, Water Resources Research, 14(2), 335-344.

[28] Panu, U. S., and T. E. Unny (1980), *Stochastic synthesis of hydrologic data based on concepts of pattern recognition: II. Application of natural watersheds*, Journal of Hydrology, 46(34), 197-217.

[29] Rogerson PA. (2001). *Statistical Methods for Geography*, SAGE Publications: London; 236 pp.

[30] Schnabel, R. R., J. B. Urban, and W. J. Gburek (1993), *Hydrologic Controls in Nitrate, Sulfate, and Chloride Concentrations*, J Environ Qual, 22(3), 589-596.

[31] Sharma, A. (2000), *Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 A strategy for system predictor identification*, J Hydrol, 239(14), 232-239.

[32] Seeger, M. et al. (2004), *Catchment soil moisture and rainfall characteristics as determinant factors for discharge/suspended sediment hysteretic loops in a small headwater catchment in the Spanish Pyrenees*, J Hydrol, 288(3-4): 299-311.

[33] Snelder, T.H. et al. (2009), *Predictive mapping of the natural flow regimes of France*, J Hydrol, 373(1-2): 57-67.

[34] Toth, E. (2009), *Classification of hydro-meteorological conditions and multiple artificial neural networks for streamflow forecasting*, Hydrol. Earth Syst. Sci., 13, 1555-1566.

[35] Toth, E. (2013), *Catchment classification based on characterisation of streamflow and precipitation time series*, Hydrol. Earth Syst. Sci., 17(3), 1149-1159.

[36] Vongvixay, A. (2012). *Mesure et analyse de la dynamique temporelle des flux solides dans les petits bassins versants*, INSA de Rennes, Rennes, 205 pp.

[37] Williams, G.P. (1989), *Sediment concentration versus water discharge during single hydrologic events in rivers*, J Hydrol, 111(1-4): 89-106.

[38] Wong, H., and B. Q. Hu (2013), *Application of interval clustering approach to water quality evaluation*, J Hydrol, 491, 1-12.

Appendices

A Running Latent Dirichlet Allocation on 3 simple flood documents

Let us consider a set of hydrological words defined as: w_1="Low value for variable 'Rainfall' on second day after the flood", w_2="High value for variable 'DOC concentration' on the day before the flood", w_3="Low value for variable 'Temperature' on fourth day after the flood", w_4="Medium value for variable 'Sulfate concentration' on the flood day".

Given a collection of 3 flood documents defined as (variables are given in Table 3):

$$D_1 = \{w_1, w_3\}, D_2 = \{w_2, w_4\}, D_3 = \{w_1, w_2, w_3, w_4\}. \quad (1)$$

We run Latent Dirichlet Allocation (LDA) asking for $K = 2$ topics. After optimizing on parameters $heta$, ϕ and z, the algorithm returns:
$\phi_1 = [0.45, 0.05, 0.45, 0.05]$: pattern corresponding to words w_1 and w_3
$\phi_2 = [0.05, 0.45, 0.05, 0.45]$: pattern corresponding to words w_2 and w_4.

Here, residuals corresponding to words w_2 and w_4 in pattern #1 and words w_1 and w_3 in pattern #2 come from prior knowledge (input parameter β, set to 0.1 for each word in each topic) used in the model.

The same is true for θ, for which sparsity is controlled by parameter α, set to 0.01 for each pattern in each document:
$\theta_1 = [0.99, 0.01]$: document #1 is explained mostly by pattern #1
$\theta_2 = [0.01, 0.99]$: document #2 is explained mostly by pattern #2
$\theta_3 = [0.50, 0.50]$: document #3 is explained equally by both patterns.

The θ parameters are closely related to the following topic assignments for observations:
$z_{1,1} = 1$: "first observation in document #1 is generated by pattern #1"
$z_{1,2} = 1$: "second observation in document #1 is generated by pattern #1"
$z_{2,1} = 2$: "first observation in document #2 is generated by pattern #2"
$z_{2,2} = 2$: "second observation in document #2 is generated by pattern #2"
$z_{3,1} = 1$: "first observation in document #3 is generated by pattern #1"
$z_{3,2} = 2$: "second observation in document #3 is generated by pattern #2"
$z_{3,3} = 1$: "third observation in document #3 is generated by pattern #1"
$z_{3,4} = 2$: "fourth observation in document #3 is generated by pattern #2".