Wavelet-Based Automated Localization and Classification of Flood Events in Hydrologic Time Series

Marco Pellegrini, Francesca Sini and Angela Celeste Taramasso

Abstract—This paper combines discrete wavelet transform (DWT) with an artificial intelligence algorithm in order to develop a new unsupervised method for fast detecting, localizing and classifying flood events in real stage-discharge time series. The proposed method does not require any a priori information such as catchment characteristics or alert flood thresholds; moreover, the simplicity of the method allows for nearly real-time execution.

Index Terms—Discrete wavelet transform, artificial intelligence, stage-discharge time series, flood event detection.

NOMENCLATURE

AI Artificial intelligence
CWT Continuous wavelet transform
DWT Discrete wavelet transform
FFT Fast Fourier transform
FWT Fast wavelet transform
IDWT Inverse discrete wavelet transform
MAD Median absolute deviation
MRA Multiresolution analysis
QMFs Quadrature mirror filters

I. INTRODUCTION

Analyzing several flood events occurred within a relevant time interval is mandatory to understand the hydrological behaviour of an instrumented catchment and to obtain a significant calibration of the basin model. The aim of this paper is the presentation of an unsupervised algorithm for detecting and classifying flood events from real-world stage-discharge time series.

The discharge time-trend of a water stream can be seen as a series of flood events ("spikes") over an almost constant base-flow. Wavelet analysis is becoming a common tool for analyzing localized signal variations within a time series and has been used in most different scientific disciplines such as speech, image and video compression and denoising [1], adaptive signal processing [2], electrocardiogram (ECG) features extraction [3] and implementation in real-time medical monitoring systems [4], spike detection in noisy neural recordings [5], weak radar signal detection and localization in noise environments [6], cosmic-ray spikes localization in radiometric measurements in the far infrared and millimeter wave spectral region [7] and numerous studies in geophysics including atmospheric cold fronts [8] and tropical convection [9]. Moreover, in combination with artificial neural networks, wavelet analysis has been applied for time series analysis [10] and prediction [11].

The proposed algorithm quickly provides the exact localization of the flood events within a discharge time series classifying the flood peaks as belonging to the rising or falling limb of the streamflow hydrograph [12]. The localization is performed through a simple artificial intelligence (AI) algorithm initialized by the position of the highest DWT coefficients.

The paper is organized as follows. Section II describes the dataset used for the examples. Section III-A introduces the concept of multiresolution analysis. This includes a discussion about the capability of such analysis to separate the frequency content of the data preserving the information in the time domain. Section III-B provides a mathematical background of the continuous wavelet transform (CWT) while Section III-C, III-D and III-E describe the wavelet analysis using a discrete notation leading to define the discrete wavelet transform. Section III-F reports a commonly used method to threshold the DWT coefficients. Section IV is about the exact localization and classification of flood peaks detected by means of DWT; the localization is performed by a local-search algorithm called hill-climbing. The effectiveness of the proposed method has been tested by applying it to discharge data derived from measured water-level time series: Section V evaluates performance of the method. Concluding remarks are given in Section VI.

II. DATA

Marche region, located in east-central Italy between the Apennines chain and the Adriatic Sea, extends over an area of about 10000 km$^2$; most of the territory is mountainous or hilly from the internal boundary towards the sea. The hilly area covers two-thirds of the region and is interrupted by wide gullies with numerous short rivers and by alluvial plains perpendicular to the Apennines chain. The parallel mountain chains contain deep river gorges.

Considering the regional hydrology, geomorphology and the vulnerability of the territory due to the high percentage of...
urbanized and industrial coastal areas, it is fundamental to provide an efficient alert system, beside a sustainable land-use planning and management, in order to reduce the inundation risk.

Since 2000 the Marche region Administration, through its Local Security and Civil Protection Department, has financed the development and the implementation of a meteorological-hydrological monitoring network (SIRMIP, Regional Meteorological-Hydrological Information System) in compliance with the Italian National Law 267/1998 (Emergency measures for the hydrogeological risk prevention and in favour of Campania region areas affected by disastrous landslide events) in order to gather, process and distribute environmental parameters such as air temperature, pressure and moisture, wind speed and direction, solar irradiance, precipitation and water level.

Marche region Multi-risk Functional Centre, within the Italian National network of Functional Centres, is the operative structure whose aim, among others, is to manage the real time monitoring network supporting the decision makers of the meteorological-hydrological alert system.

Regarding the stream flow monitoring, SIRMIP network consists to this day of 76 non-contact ultrasonic water-level sensors engineered by LIF company [13] having a resolution of 1 cm and an accuracy of ±1 cm; in order to reduce the systematic errors in such devices [14], sensors are provided with an internal temperature probe and a firmware algorithm that automatically compensate for the temperature-related variations in the ultrasonic wave propagation speed. Fig. 1 shows Marche region with the localization of the water-level sensors.

For several stream sections where level-sensors are installed, the parameters of the stage-discharge relation (also known as rating curve) assumed as logarithmic [15] have been determined:

$$Q = p(h - e)^\beta$$

(1)

where $Q$ is the discharge and $h$ is the gauge height of the water surface, $e$ is the effective gauge height of zero flow, $(h - e)$ is the effective depth of water on the cross-section of the stream, $\beta$ represents the slope of the rating curve and $p$ is a constant such that $p = Q$ when $(h - e) = 1$.

Fig. 2 shows the water level measured at Aspio Terme section, few kilometers south of Ancona, during year 2008 (17568 data considering a sampling interval of half an hour). Aspio river watershed lies in a relatively small drainage basin of about 160 km$^2$ whereof one half upstream of Aspio Terme section. In the following this time series will be taken as an example for illustrating the algorithm.

Stage (and discharge for the stream sections provided with a rating curve) data used in this paper are available in digital form from the SIRMIP on-line web site [16].

III. WAVELET ANALYSIS

A. Multiresolution analysis

Wavelet analysis consists of decomposing a signal or an image into a hierarchical set of approximations and details; among other things it is a useful tool for detecting transient signals within a time series [17].

Taking advantage of the fact that a spike spreads all over the frequencies while the underlying signal is limited to a narrow band, it is possible to separate the frequency content of the data without losing the information in the time domain by decomposing the signal into a time-scale space; this approach is known as multiresolution analysis (MRA) [18].

MRA is designed to give good time resolution and poor frequency resolution at high frequencies and good frequency resolution and poor time resolution at low frequencies. MRA approach makes sense especially when the signal at hand has high frequency components for short periods of time and low frequency components for long periods; many natural signals encountered in practical applications, as a flood event within an hydrograph, are often of this type. A peculiarity of MRA is the ability to perform local analysis detecting signal patterns that are barely visible too.

Fig. 3 shows the time-scale representation provided by MRA: as one needs to look at the signal at smaller scales, MRA allows for more localized analysis and vice versa [19]; the boxes are known as the Heisenberg boxes due to the Heisenberg uncertainty principle which tells us the minimum area (the product of the time and frequency spreads) that boxes can have.

MRA does not use a time-frequency representation, but rather a time-scale resolution region; more details about the connection between scale and frequency can be found in [20].
Fig. 3. Time-scale representation provided by multiresolution analysis [1]. A wavelet orthonormal basis decomposes the time axis in dyadic intervals whose sizes have an exponential growth; at each scale, the area of the scale-time boxes holds steady (see Section III-C).

Fig. 4. Estimated Aspio Terme discharge time series during year 2008 (upper side) and highlights two discharge spikes occurring at each zoom selection areas related to water-volume flood peaks differing one another by one order of magnitude (lower side). Time indexes 6720 and 14820 correspond to May 20 at midnight and November 4, 2008 at 6:00 PM local time (UTC+1) respectively.

As in the case of maps, high scales correspond to a non-detailed global view of the physical phenomenon, and low scales correspond to a detailed view. In terms of frequency, low frequencies (high scales) correspond to a global information about the signal, whereas high frequencies (low scales) correspond to a detailed information about relatively short time intervals of the signal.

Many geomorphological and geophysical phenomena such as drainage networks, erosion, earthquakes and floods exhibit scale-invariance and self-affine properties [21]; consequently, MRA methods, such as the wavelet decomposition, are becoming a common investigating tool for physical applications where multiscale phenomena occur [22]. Fig. 4 reports the estimated Aspio Terme section discharge time series for the year 2008 (upper side) and highlights two discharge spikes occurring at each zoom selection areas related to water-volume flood peaks differing one another by one order of magnitude (lower side).

B. Continuous wavelet transform

Wavelet transform provides a multiresolution representation using a set of analyzing functions derived from a base function $\psi(t)$, the mother wavelet. Mother wavelet is a function of finite energy, i.e., $\psi \in L^2(\mathbb{R})$, normalized $\|\psi\|_2 = 1$, zero average $\int_{\mathbb{R}} \psi(t) dt = 0$ and centered in the neighborhood of the origin. In most situations it is useful to restrict $\psi(t)$ to be a continuous function with a higher number $M$ of vanishing moments, i.e., for all integer $m < M$, $\int_{\mathbb{R}} t^m \psi(t) dt = 0$.

From $\psi(t)$, by translations and scaling, one can obtain a wavelet basis therefore a family of time-scale waveforms

$$\psi_{a, b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t - b}{a}\right), \quad a, b \in \mathbb{R}$$  \hspace{1cm} (2)

where $a > 0$ is the scale and $b$ the translation. The continuous wavelet transform (CWT) of a signal $s(t)$ is the projection of $s(t)$ onto the wavelet basis:

$$C_s(a, b) = \int_{\mathbb{R}} s(t) \psi_{a, b}(t) dt \hspace{1cm} (3)$$

$C_s(a, b)$ values clearly depend on the shape of the chosen mother wavelet. For mother wavelets of compact support [23], the domain of integration in (3) is limited to the support: in this way, the wavelet coefficient depends only on the portion of $s(t)$ within the support. If the scale is relatively small, a wavelet of compact support allows for local analysis detecting transient signals.

C. Discrete wavelet transform

Calculating CWT as in (3) for all possible scale and translation is a fair amount of work. Choosing only a finite and discrete subset of scales and positions based on powers of two (so-called dyadic scales and positions) then the analysis will be much more efficient and just as accurate: such an analysis is known as discrete wavelet transform (DWT). Dealing with dyadic scales means that the scale $a$ will be of the form $2^j$, $j = 1, 2, 3, ...$ and then, within a given scale, the translations will be separated by multiples of $2^j$ [24].

In 1988 I. Daubechies has constructed compactly supported dyadic orthonormal wavelet bases for $L^2(\mathbb{R})$ of arbitrarily high regularity [23]; the regularity property is determined by the number of vanishing moments of the mother wavelet. In general, Daubechies wavelets are chosen to have the highest number $M$ of vanishing moments for given support width $2M$ [25].

In 1989, Mallat proposed a fast wavelet decomposition and reconstruction algorithm (also known as fast wavelet transform, FWT) [18], schematized as in Fig. 5 (decomposition and reconstruction are in the upper and lower side respectively). The input signal $s_N$ is a discrete $N$-samples sequence. Since we are dealing with dyadic scales, we consider $N$ as a power of two (or at least a multiple of power of two) therefore $N = 2^k$ in order the scheme to be efficient. For this reason the original Aspio Terme section data-stream has been zero-padded to $N = 2^{15}$, the smallest power of two greater than or equal to the original number of data points.

At the most basic level of the decomposition process $s_N$ passes through two complementary (half-band) filters: a low-pass filter $L$ in order to obtain $a_1$, an approximation of the signal (low frequency components), and an high-pass filter $H$ to highlight the details $d_1$ (high frequency components); output
the first level of decomposition, both of the signal spans only half of the previous frequency band, it doubles the frequency resolution since the frequency band characterizes the entire signal and, at the same time, the time resolution since only half the number of original data points in \( s_N \) can be discarded according to the Nyquist’s rule. The scheme implements a three-level decomposition and reconstruction.

For each scale, the product of the number of coefficients and their spacing in time is always \( N \); according to the Heisenberg uncertainty principle, at scale \( 2^{-j-1} \) the time location is known with a precision of plus or minus \( 2^{j-1} \) data points in \( s_N \).

The frequency domain representation of the DWT corresponding to a three-level decomposition is shown in Fig. 6. If the original signal ranges from 0 to \( f_m \), first level of decomposition \( d_1 \) will range from \( 1/2 f_m \) to \( f_m \), \( d_2 \) from \( 1/4 f_m \) to \( 1/2 f_m \) and \( d_3 \) from \( 1/8 f_m \) to \( 1/4 f_m \), low frequencies from 0 to \( 1/8 f_m \) characterize the scaling coefficients \( a_3 \).

Analogous considerations can be made for the process leading to the reconstructed signal \( s_N \), the inverse discrete wavelet transform (IDWT): where DWT involves filtering and down-sampling, IDWT consists of up-sampling and filtering. The up-sampling process inserts zeros between samples in order to lengthen the discrete signal.

D. Quadrature mirror filters

Choosing the four half-band filters \( L(z) \), \( H(z) \), \( L'(z) \) and \( H'(z) \) (in the \( z \)-domain) as a system called quadrature mirror filters (QMFs), it is possible to achieve perfect reconstruction of \( s_N \) and all filters are determined by the 2M-length low-pass filter \( L(z) \) [26]. QMF perfect reconstruction is obtained combining the symmetry constraint between the decomposition filters

\[
H(z) = -z^{-(2M-1)} L(-z^{-1})
\]

and the aliasing-cancellation conditions

\[
L'(z) = z^{-(2M-1)} L(z^{-1}) = H(-z)
\]

\[
H'(z) = z^{-(2M-1)} H(z^{-1}) = -L(-z)
\]

In the time domain, given the low-pass decomposition filter \( l(m) \) with \( m = 0, ..., 2M-1 \), \( h(m) = (-1)^m l(-m-2M+1) \), i.e., all odd-index coefficients of reversed \( l(m) \) are negated, and reconstruction filters \( l'(m) \) and \( h'(m) \) are reversed version of \( l(m) \) and \( h(m) \) respectively (alternating flip QMFs [26]).

Orthogonal wavelets, such as the Haar wavelets and related Daubechies wavelets, Coiflets, and some developed by Mallat, use QMFs. These filter banks implement a fast orthogonal wavelet transform that requires only \( O(N) \) operations for a signal of size \( N \) or, in other terms, the execution time is linear with the signal length. FWT is less computationally complex compared to the Fast Fourier Transform (FFT) taking \( O(N \log_2 N) \) time [1]: this computational advantage is not
inherent to the transform, but reflects the choice of a logarithmic division of frequency, in contrast to the equally-spaced frequency division of the FFT. Due to its simplicity, FWT can be hardware implemented allowing for real-time execution (as for FFT).

A significant difference between FFT and FWT is that in former case the algorithm needs to complete all the steps in order to obtain valid coefficients while in the latter case one can stop the decomposition at each level however being able to reconstruct the signal through IDWT.

E. Daubechies wavelets and DWT calculation

Daubechies wavelets are not defined in terms of wavelet and scaling functions; in fact, it is not possible to write them down in closed form [27]. A Daubechies wavelet family is completely characterized by the coefficients of the scaling function \( l(m) \) with \( m = 0, \ldots, 2M - 1 \). In this paper we use Daubechies 20-coefficient wavelet \( (M = 10) \) vanishing moments) [28]. Further details about the construction of the Daubechies scaling functions can be found in [1]. The core procedure used to calculate DWT and IDWT has been implemented as given in [29].

Fig. 7 shows the DWT decomposition of Aspio Terme section discharge time series up to scale \( a = 4 \) \((j = 3)\); each wavelet sequence has been linearly-spaced as in (4) in order to align the flood peaks and the DWT coefficients.

In general, most prominent frequencies in the original signal appear as high amplitudes in that region of the DWT sequence that includes those particular frequencies; as said previously, the time localization has a resolution that depends on which scale they appear. Since we are interested in detecting spike features within \( s_N \), we can focus our attention only on the smallest scale in which the time localization of the frequencies will be more precise since they are characterized by the highest number of samples. In particular, at scale \( a = 1 \) \((j = 1)\) the time location of a given frequency in the original stream can be determined starting from DWT with a precision of plus or minus 1 data point.

F. Wavelet thresholding

By applying DWT we obtain a multiscale representation of a signal in terms of its wavelet coefficients; wavelet transform provides a sparse representation of the signal [1] where only a few coefficients are significantly different from zero. This can be noticed in Fig. 7 where at the smallest scale (strip \( d_1 \)) the effect of flood events on the DWT coefficients is clearly visible. For purpose of unsupervised flood-event detection, we have to separate these non-zero coefficients from the other coefficients by imposing a threshold calculated from the sampled data.

To obtain the threshold, we borrow ideas from Donoho and Johnstone [30] who studied the problem of nonlinear estimation of signals under a sparse representation. Their wavelet thresholding method is based on accepting only those coefficients that exceed a given threshold. For a near-optimal performance it is sufficient to choose the commonly used universal global threshold defined as \( T = \sigma_B \sqrt{2 \log N} \), where \( \sigma_B \) represents the estimate of the background signal standard deviation and, as before, \( N \) is the number of samples of the time series.

Here we investigate the utilization of an adaptive thresholding rule as proposed by [31], which amounts to using a threshold that depends on the resolution level of the wavelet transform. In other words at scale-index \( j \) we have

\[
T_j = \sigma_B j \sqrt{2 \log N}
\]

where \( \sigma_B = \sqrt{2} \ M A D_j/0.6745 \), computed by means of the median absolute deviation (MAD) of the wavelet coefficients at level \( j \), i.e., \( M A D_j = \text{median}(|d_j - \text{median}(d_j)|) \).

Denoting with \( n_T \) the set of indexes related to the coefficients in \( d_1 \) exceeding the threshold \( T_1 \), i.e.,

\[
|d_1(n)| > T_1 \quad \forall n \in n_T
\]

through (4) it is possible to obtain \( \tau_1(n_T) \), a coarse localization of flood peaks in the original time series.

IV. PEAK LOCALIZATION AND EVENT CLASSIFICATION

The exact time-localization of flood peaks is performed by means of a hill-climbing search algorithm [32], one of the most basic local-search techniques; in Fig. 8 the pseudo-code of the algorithm is shown.
The initial-state of the algorithm is assumed to be $\tau_1(n_T)$ and a generic node is the set of time-indexes approaching local maxima in the original data-stream; at each step the current node is replaced by the best neighbor according to the discharge value. Since the time localization of a given frequency is determined from sequence $d_1$ with a precision of one data point in $s_N$ (see Section III-C), the convergence of hill-climbing algorithm is very fast.

Once the peak of a generic flood event $f$ has been localized, in order to define its duration $T_f$ we considered the monthly-averaged flow $Q_m(t)$ as a threshold; in this way we have taken into account the temporal variability of flood characteristics and causes [33]. The intersections between $Q_m(t)$ and discharge $Q(t)$ define the start and end point of $f$:

$$Q(t) \geq Q_m(t) \quad \forall t \in T_f$$ (9)

One of the flood events occurred within the Aspio Terme time series and the threshold $Q_m(t)$ are shown in Fig. 9: the peak of 29 m$^3$s$^{-1}$ has been correctly localized by the proposed method at index 16647, corresponding to December 12, 2008 at 7:30 PM local time.

The event in Fig. 9 is composed by two sub-events, one with a peak of 14 m$^3$s$^{-1}$ at index 16625 and the other including the highest discharge value of the whole event. These two sub-events were detected and localized by the algorithm as distinct events; since they are still part of the same event sharing the same threshold crossing, we considered a post-processing phase within the proposed method in order to classify all the sub-events regarding the main peak. In particular, the algorithm classifies a sub-event as belonging to the rising or falling limb of the streamflow hydrograph also identifying the lowest discharge value between two sub-events (a minimum of 3 m$^3$s$^{-1}$ at index 16635 can be observed in Fig. 9). The algorithm provided the localization and the classification of 32 main events during whole year 2008 for the Aspio Terme section time series.

A different choice regarding the stream discharge threshold will obviously lead to a different duration $T_f$ and then a probably different sub-event classification: since the event classification does concern neither the DWT calculation nor the wavelet thresholding but only the post-processing phase of the algorithm, the flow threshold choice does not influence the performance of the proposed method. For nearly real-time monitoring purposes or for selecting significant flood events to calibrate an hydrological model, any available a priori information about the observed streamflow, just as a known alert flow level threshold, can be easily included in the algorithm in order to ignore negligible event features.

The applied wavelet approach can also be useful for pointing out the presence of relevant dam releases upstream of the analyzed cross-section, by detecting temporal flood features that could be compared to a release schedule. As an example, in Fig. 10 we report the Brecciarolo hydrograph for the year 2008; in this time series are clearly visible few main events superimposed to a noisy background signal. Brecciarolo water-level sensor is located on Tronto river, south of Marche region.

One-stage DWT decomposition of Brecciarolo discharge time series using Daubechies 20-coefficient wavelet. As for Aspio Terme section, the original data-stream has been zero-padded to $2^{15}$ samples (bottom). $d_1$ are the wavelet coefficients (center) and $a_1$ the scaling coefficients (top). On the right is shown the length of each strip.
detected events has its maximum discharge value of 19 m$^3$s$^{-1}$ at index 1182 corresponding to January 25, 2008 at 3:00 PM local time. Qualitatively speaking, it is easy to recognize a regular flood-peak cadence having reference to a typical dam release schedule.

V. PERFORMANCE EVALUATION

The method described in the previous sections has been tested exploiting it to automatically detect, localize and classify flood events from year-long discharge data time series gathered from the SIRMIP network database. Data coming from 10 stream sections provided with a rating curve have been analyzed from year 2005 to 2008.

As mentioned in Section III-D, DWT computational time is linear with the signal length while the execution time of the classification phase of the algorithm depends on the number of flood peaks localized within the data-stream after the wavelet thresholding.

Running the algorithm on a commercial PC with a 3.4 GHz CPU and 1 GB of RAM, it is possible to process one-year data in less than 9 s (worst case) with an average time of about 4 s. This result makes the analysis of hydrological data time series much faster.

VI. CONCLUSION

The proposed algorithm has been extensively tested over the available Marche region dataset allowing for an unsupervised localization and classification of flood events occurred within a discharge time series. The classified flood events can be so easily exploited for hydrological purposes and watershed basin modelling, with a relevant reduction in data analysis time.

Without any a priori information, the event classification has been developed defining the duration of an event by means of the monthly-averaged flow as a threshold. The classification phase of the algorithm can be further improved including any available temporal information about the flood event duration (such as the watershed time of concentration [12] or the average duration of investigated historical flood events) in order to better distinguish the flood sub-events at the analyzed cross-section.

ACKNOWLEDGMENT

The authors would like to acknowledge the anonymous reviewers and the editorial staff. This research has been partially funded by Dipartimento per le Politiche Integrate di Sicurezza e per la Protezione Civile, Regione Marche, and Dipartimento di Informatica Sistemistica e Telematica (DIST), Università degli Studi di Genova.

REFERENCES

Marco Pellegrini received the laurea degree in telecommunications engineering from the University of Florence, Florence, Italy, in 2001 and the Ph.D. degree in methods and technologies for environmental monitoring from the University of Basilicata, Potenza, Italy, in 2006. From 2002 to 2005 he was with the Department of Electronics and Telecommunications, University of Florence, Italy and with the Institute of Applied Physics "Nello Carrara", Italian National Research Council, Florence, Italy. Since 2007 he has been with LIF company, Florence, Italy. His research interests include signal and data processing for passive and active remote sensing, multiresolution analysis and applications of wavelets to geophysical signals.

Francesca Sini received the laurea degree in environmental engineering from La Sapienza University, Rome, Italy, in 2001 and the Ph.D. degree in Methods and Technologies for Environmental Monitoring from the University of Basilicata, Potenza, Italy, in 2006. From 2002 to 2005 she joined CIMA, University of Genova, Italy. In 2004 she was a visiting student with the Ralph M. Parsons Laboratory for Environmental Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. Since 2007 she has been with the Functional Centre for Meteorology and Hydrology of Regione Marche, Italy. Her research interests include catchment-scale hydrological processes, flood alert systems and civil protection, land-atmosphere interaction and data assimilation.

Angela Celeste Taramasso graduated with honors in mechanical engineering at the University of Genoa, Italy, in 1983. She received the Ph.D. degree in hydrodynamics from the University of Padua, Italy, in 1989, and her thesis was honored with the Filippo Arredi award. Since 1990 she has been an Associate Researcher at University of Genoa, Italy, now with the Department of Communication, Computer and System Sciences (DISt). She has been involved in several national and international research programs on environmental engineering and natural hazards. Her research interests include hydraulic and hydrological modeling, coastal engineering and civil protection planning and risk management.