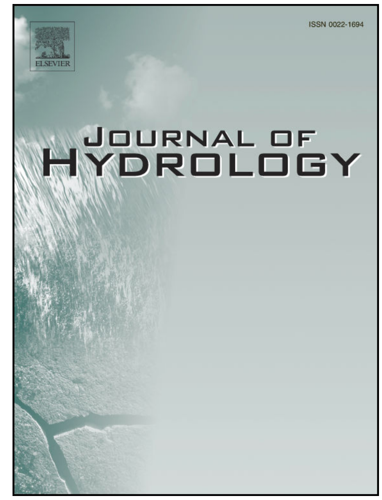
Journal Pre-proofs



Research papers

Multivariate Event Time Series Analysis using Hydrological and Suspended Sediment Data

Ali Javed, Scott D. Hamshaw, Donna M. Rizzo, Byung Suk Lee

PII: S0022-1694(20)31263-4
DOI: https://doi.org/10.1016/j.jhydrol.2020.125802
Reference: HYDROL 125802

To appear in: *Journal of Hydrology*

Received Date: 2 July 2020
Revised Date: 22 November 2020
Accepted Date: 23 November 2020

Multivariate Event Time Series Analysis using Hydrological and Suspended Sediment Data

Ali Javed[*,1], Scott D. Hamshaw[2], Donna M. Rizzo[2], and Byung Suk Lee[1]

[1]Department of Computer Science, University of Vermont, Burlington, VT, USA

[2]Department of Civil & Environmental Engineering, University of Vermont, Burlington, VT, USA

[*]Corresponding author email: ali.javed@uvm.edu

[*]Mailing Address: 82 University Place, Innovation STEM Building, Burlington, VT 05405, United States

Highlights

- New analysis for clustering multivariate time series for hydrological events
- Time series clustering and 2-D hysteresis provide insight into storm event behavior
- The new clustering approach is applied to Vermont watershed data
- Synthetically generated C-Q data provide effective environmental model validation

Abstract

Hydrological storm events are a primary driver for transporting water quality constituents such as suspended sediments and nutrients. Analyzing the concentration (C) of these water quality constituents in response to river discharge (Q), particularly when monitored at high temporal resolution during a hydrological event, helps to characterize the dynamics and flux of such constituents. A conventional approach to storm event analysis is to reduce C-Q time series to two-dimensional (2-D) hysteresis loops and analyze these 2-D patterns. While informative, this hysteresis loop approach has limitations because projecting the C-Q time series onto a 2-D plane obscures detail (e.g., temporal variation) associated with the C-Q relationships. In this paper, we address this limitation using a multivariate event time series (METS) clustering approach that is validated using synthetically generated event times series. The METS clustering is then applied to river discharge and suspended sediment data (acquired through turbidity-based monitoring) from six watersheds in the Lake Champlain Basin located in the northeastern United States, and results in identifying four common types of hydrological water quality events. Statistical analysis on the events partitioned by both methods (METS clustering and 2-D hysteresis classification) helped identify hydrometeorlogical features of common event types. In addition, the METS and hysteresis analysis were simultaneously applied to a regional Vermont dataset to highlight the complimentary nature of using them in tandem for hydrological event analysis.

Keywords: event analysis, streamflow, suspended sediment, clustering, multivariate time series, water quality sensors

1 Introduction

Characterizing the processes associated with rainfall-runoff events is an essential part of watershed research; and studying the dynamics that drive these processes (e.g., the timing and location of water quality constituent fluxes through the landscape) has many applications in the hydrological sciences. These include identifying sources of erosion present in a watershed (Sherriff et al., 2016), monitoring for shifts in watershed function (Burt et al., 2015), improving hydrological model forecasts (Ehret and Zehe, 2011), and informing watershed conservation and management efforts (Bende-Michl et al., 2013; Chen et al., 2017). Environmental managers and scientists often analyze hydrological data (e.g., suspended sediment concentration and streamflow) at an event scale — in this work, the period of storm-runoff resulting from a rainfall event – because this period is the primary mechanism for transporting many constituents of concern (Dupas et al., 2015; Sherriff et al., 2016). The timing of constituent delivery relative to stream discharge is complex and often exhibits a high degree of variability, especially when the monitoring frequency is high (Minaudo et al., 2017); and unsurprisingly, the relationship between multiple responses during a single event (e.g., discharge and water quality constituents) is often not linear (Onderka et al., 2012). However, despite the inherent complexity and dynamic behavior, the analysis of concentration-discharge (C-Q) relationships to infer mechanistic watershed processes at the event scale has a long tradition in hydrology, geomorphology and ecology (Aguilera and Melack, 2018; Burns et al., 2019; Williams et al., 2018;

Malutta et al., 2020).

A fundamental feature of suspended sediment and solute transport in rivers is that the concentration of such constituents is often not in phase with the associated stream discharge, resulting in hysteresis being observed in the C-Q relationship. Williams (1989) was one of the first to use hysteresis patterns to study hydrological storm events, identifying six classes of hydrological events and offering linkages between the hysteresis classes and watershed processes. While the study focused on suspended sediment concentration (SSC) data, these event classifications have been widely adopted in studies of both sediment and solutes, and continue to be used today to group storm events (e.g., Aguilera and Melack, 2018; Rose et al., 2018; Keesstra et al., 2019). An alternate to using 2D hysteresis patterns for categorization is to simplify the C-Q relationship into a scalar hysteresis index (Lloyd et al., 2016b). While both approaches are effective for inferring certain physical processes, each loses some information associated with the raw time series data, because both approaches "collapse" the time dimension, either by projecting the C-Q data onto a two-dimensional plane, or reducing the information into a scalar value (an index). Thus, temporal information associated with the original times series, such as the rate of change of a variable as well as aspects of its shape (e.g., linear, convex, concave), may be lost. With the increasing availability of high frequency sensors and associated data processing tools, it is now possible to leverage the temporal information embedded in multiple time series and fuse the data with complementary event analysis schemes such as hysteresis loop classification (Williams, 1989).

A few hydrological studies have used univariate time series (e.g., discharge) to quantify the similarity between storm events for forecasting purposes. Ehret and Zehe (2011) used manual feature extraction to propose a similarity measure for discharge time series that leverages hydrograph attributes such as the rising limb, peak and receding limb. Such manual feature extraction works well for hydrographs, but may not generalize to multivariate water quality time series. Ewen (2011) modified the minimal variance matching algorithm (Latecki et al., 2005) to quantify the similarity between two hydrographs. Presented with a hydrograph defined by a sequence of discharge measurements (called a "query sequence"), the method finds a target hydrograph that contains a sub-sequence most similar to the query sequence. Because only a portion of the target sequence is matched (Latecki et al., 2005), similarity is not symmetric in both directions (i.e., $d(x, y)! = d(y, x)$) and, hence, may not be appropriate for use in clustering hydrological event data. Wendi et al. (2019) used recurrence quantification analysis and cross-recurrence plots to measure similarity between recurring hydrograph patterns. Recurrence quantification analysis is useful for large flood events (particularly those with multiple peaks); however, when the events are delineated, as is done in our work, the approach may not be appropriate. Regardless, none of the above classification methods were designed for analyzing events with multivariate time series.

Several studies have clustered storm events using event metrics and/or coefficients of best fit models. Dupas et al. (2015) used dynamic time warping (DTW) and K-means clustering to cluster re-scaled time series of phosphorus concentration. They manually select an ideal hydrograph and use the DTW algorithm to align each hydrograph in the dataset to the ideal hydrograph. Using these aligned hydrographs, the respective event phosphorus concentration graphs are then clustered to find dominant response patterns associated with physical processes occurring in the watershed. Bende-Michl et al. (2013) used high frequency data to build a database of events summarized by metrics such as precipitation, discharge, runoff coefficient and maximum discharge. These metrics were then used in cluster analysis to study nutrient dynamics in the Duck River, in north-western Tasmania, Australia. Minaudo et al. (2017) applied the non-linear empirical modeling method of Mather and Johnson (2014) using continuous records of turbidity and discharge to estimate high frequency phosphorus concentration values from low frequency (e.g., weekly) sampling. They then clustered storm events using sets of model coefficients that were fit to each storm event. The coefficients were re-calibrated for each cluster to obtain one set of coefficients representative of all storm events in the cluster. Mather and Johnson (2015) modeled event turbidity as a function of event discharge using a power-law model, and performed cluster analysis on the model parameters to select the number of hysteresis loop categories, thereby avoiding *a priori* selection of the number of classes. While all of these works extract event information from two monitored variables (e.g., C and Q), none directly use the full time series (i.e., without transformation or feature extraction) associated with both variables to cluster storm events.

In this paper, we present a data-driven approach for clustering multivariate water quality time series at the event scale. We refer to this method as METS (multivariate event time series) clustering throughout the remainder of the manuscript; and show proof-of-concept using two variables: concentration (C) and discharge (Q). These time series may be visualized as trajectories in a 3-D space, namely a C-Q-T plot. Our concentration data comprise three years of high-resolution riverine suspended-sediment concentration (SSC) time series – for generalizability, referred to simply as C – collected from six watershed sites in Vermont. The efficacy of the approach is demonstrated both qualitatively, using multi-dimensional visualizations (i.e., C-Q-T plots), and quantitatively using metrics that summarize event characteristics. We also highlight the complementary nature of using METS in tandem with other analysis schemes, in this work – the C-Q hysteresis patterns of Williams (1989).

2 Study Area and Data

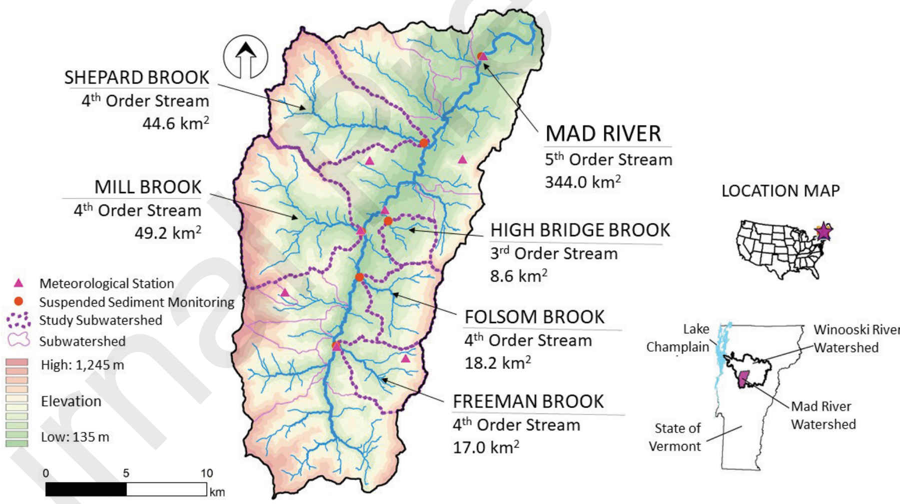


Figure 1: The Mad River watershed and study sub-watersheds within the Lake Champlain Basin of Vermont.

Our study area, located in the Mad River watershed (Figure 1) in the Lake Champlain Basin and central Green Mountains of Vermont, is the site of several ongoing geomorphic and sediment dynamics studies at the University of Vermont (Stryker et al., 2017; Wemple et al., 2017; Hamshaw et al., 2018). Continuous streamflow and suspended sediment monitoring data (SSC) were collected for more than 600 storm events in this watershed (and its five sub-watersehds) between October 19^{th}, 2012 to August 21^{th}, 2016 (Table 1). Hamshaw et al. (2018) used this dataset to automate and demonstrate possible refinements to the 2D (C-Q) hysteresis classifications of Williams (1989). Turbidity data were collected every 15 minutes using turbidity sensors and SSC-turbidity regression models were used to calculate SSC (see Hamshaw et al. (2018) for details). Discharge data were obtained from the United States Geological Survey (USGS) stream gauges or calculated using stage-discharge rating curves. The individual storm events were extracted from the continuous sensor records using a semi-automated approach based on thresholds to detect events and manual identification of storm end points. Meteorological data (rainfall and soil moisture) were also collected over the monitoring period and summarized into 24 storm event metrics (see Table 2); for full details on data collection and event delineation methodology, readers are referred to Hamshaw et al. (2018).

Table 1: Number of storm events and monitoring start and end dates for each watershed study site.

Site	**Number of events monitored**	**Monitoring start date**	**Monitoring end date**
Freeman Brook	54	Jun 2^{nd}, 2013	Nov 17^{th}, 2013
Folsom Brook	96	Jul 17^{th}, 2013	Sept 13^{th}, 2015
Mill Brook	158	Oct 19^{th}, 2012	Dec 23^{rd}, 2015
High Bridge Brook	41	Jun 6^{th}, 2013	Nov 17^{th}, 2013
Shepard Brook	106	Jul 18^{th}, 2013	Dec 23^{rd}, 2015
Mad River (main stem)	148	Oct 29^{th}, 2012	Aug 21^{th}, 2016
All Sites	**603**	**Oct 19^{th}, 2012**	**Aug 21^{th}, 2016**

The Mad River watershed ranges in elevation from 132 m to 1,245 m above sea level and is predominantly forested except for the valley bottom, which features agriculture, village centers, and other developed lands (Supporting Information Table S1). The watershed has a mean annual precipitation ranging from approximately 1,100 mm along the valley floor to 1,500 mm along the upper watershed slopes (PRISM, 2019). Soils range from fine sandy loams derived from glacial till deposits in the uplands to silty loams from glacial lacustrine deposits in the lowlands. Erosional watershed processes include bank erosion, agricultural runoff, unpaved road erosion, urban storm water, and hillslope erosion. Similar to many watersheds in Vermont, reducing excessive erosion and sediment transport in the Mad River is a focus of several management efforts including stormwater management practices, streambank stabilization and river conservation.

Table 2: Description of the 24 storm event metrics used in this work.

Metric	Description
Hydrograph/ Sedigraph characteristics	
T_Q	Time to peak discharge (hr)
T_{SSC}	Time to peak TSS (hr)
T_{QSSC}	Time between peak SSC and peak flow (hr)
Q_{Recess}	Difference in discharge value at the beginning and end of event
SSC_{Recess}	Difference in concentration value at the beginning and end of event
D_Q	Duration of stormflow (hr)
FI	Flood intensity
SSC_{Peak}	Peak SSC (mg/L)
HI	Hysteresis index
Antecedent conditions	
T_{LASTP}	Time since last event (hr)
A3P	3-Day antecedent precipitation (mm)
A14P	14-Day antecedent precipitation (mm)
$SM_{SHALLOW}$	Antecedent soil moisture at 10 cm depth (%)
SM_{DEEP}	Antecedent soil moisture at 50 cm depth (%)
BF_{NORM}	Drainage area normalized pre-storm baseline flow($m^3/s/km^2$)
Rainfall characteristics	
P	Total event precipitation (mm)
P_{max}	Maximum rainfall intensity (mm)
D_P	Duration of precipitation (hr)
T_{PSSC}	Time between peak SSC and rainfall center of mass (hr)
Streamflow and sediment characteristics	
BL	Basin lag
Q_{NORM}	Drainage area normalized stormflow ($m^3/s/km^2$)
Log(Q_{NORM})	Log-normal stormflow quantile (%)
SSL_{NORM}	Drainage area normalized total sediment (kg/m^2)
$FLUX_{NORM}$	Drainage area and flow normalized sediment flux ($kg/m^3/km^2$)

In addition to the Mad River watershed sites, we created an expanded regional dataset by adding 190 events from three additional watersheds (Hungerford Brook, Allen Brook, and Wade Brook) in the Lake Champlain Basin to the existing ($n = 603$) Mad River events, and another 21 events from within the Mad River watershed during the period from April $3^{rd}, 2007$ to November $25^{th}, 2016$. This results in a total of 814 storm events from nine watersheds, hereafter referred to as the "regional Vermont dataset". Hungerford Brook, Allen Brook, and Wade Brook are watersheds with ongoing monitoring efforts (Vaughan et al., 2017) that represent a spectrum of land uses (e.g., agricultural, forested, and developed, respectively) and feature varied topographic characteristics (Supporting Information Table S1). Data from these sites, and supplemental events from the Mad River do not have the corresponding hydrometeorological data metrics associated with the Mad River dataset and thus were not the focus of our primary analyses.

3 Methods

3.1 Event Time Series Processing

The sensor data collected during individual storm events are conceptualized as trajectories and may comprise *multivariate* time series of two or more variables. For example, two (univariate) time series, $TS1 = \langle V1_1, V1_2, V1_3, ..., V1_n \rangle$ and $TS2 = \langle V2_1, V2_2, V2_3 ..., V2_n \rangle$, when combined, make a bivariate time series $\mathbf{TS} = \langle (V1_1, V2_1), (V1_2, V2_2), ..., (V1_n, V2_n) \rangle$. This approach can be generalized to the multivariate case of a matrix of m variables and n time steps (Supporting Information Figure S1).

The time series in this work (discharge and SSC) were collected *in situ* using multiple environmental sensors. These data typically contain noise, have missing values, and often require pre-processing (i.e., filtering) to extract general trends in the C-Q relationship. In addition, because of our interest in comparing C-Q relationships across hydrological events, we normalized both the length of the time series as well as the magnitude of each variable individually over each event (Figure 2), as is commonly done in C-Q analyses. Pre-processing steps were performed as follows:

Smoothing: To reduce noise, the discharge and concentration time series were smoothed using the

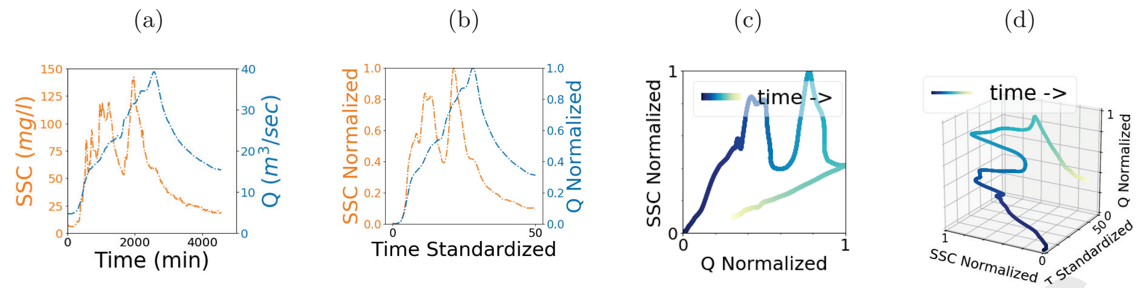


Figure 2: Pre-processing of (a) raw C and Q time series, (b) smoothed and normalized C and Q time series, and the resulting (c) C-Q plot, and (d) C-Q-T plot for an individual (delineated) storm event.

Savitsky-Golay Filter (Savitzky and Golay, 1964). We selected a third-order, 21-step filter for the Mad River (main stem) and a fourth-order, 13-step filter for each of the five sub-watersheds. To preserve the peaks and overall shape of the event data, the filter order and step size were selected based on visual inspection of the resulting event time series in a manner similar to Hamshaw et al. (2018).

Standardization of event length: Discharge and concentration time series were re-scaled to a uniform length of 50 time steps for all events using univariate spline fitting (Dierckx, 1993). The number 50 was selected empirically as the minimum number of data points that preserves the shape and characteristics of the event time series. Standardizing all events to have the same length ensured that clustering was not affected by the duration of the event but by the relative rate of change of C-Q variables. We note that this re-sampling was performed separately from the calculation involving event metrics (Table 2) based on the original data.

Normalization of magnitude: The discharge and concentration time series were scaled individually to values between 0 and 1. This ensured that the clustering is not affected by the magnitude of the individual time series but by the orientation of change (e.g., clockwise and counter-clockwise), and the shape (e.g., linear, convex and concave). Normalizing the magnitude of variables is common for a meaningful comparison between time series (Rakthanmanon et al., 2012).

3.2 Concentration-discharge (C-Q) Hysteresis Classification

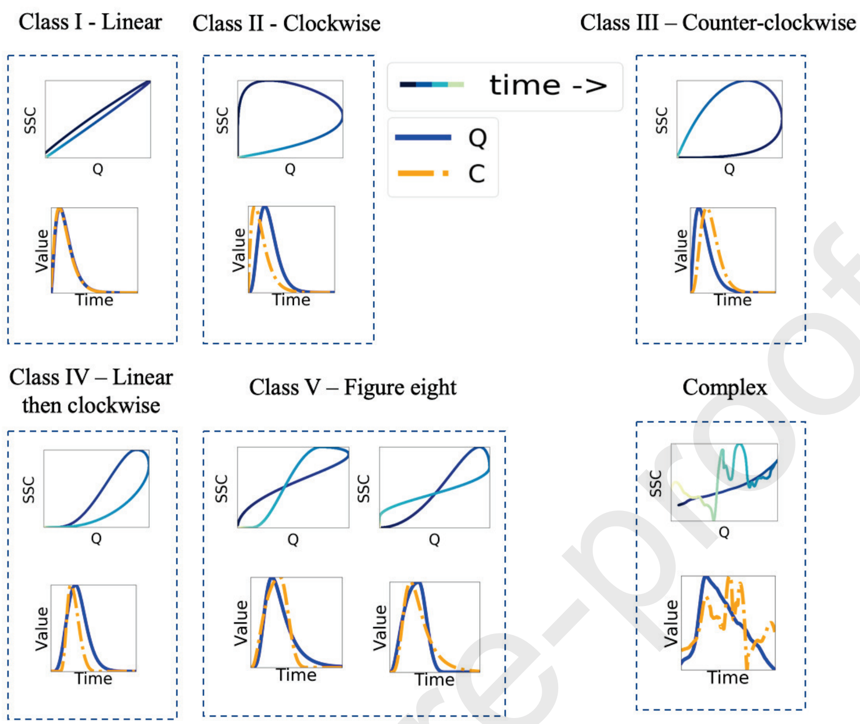


Figure 3: Six class scheme for concentration-discharge hysteresis loops (top panels) and corresponding hydrographs and sedigraphs (lower panels, solid and dot-dashed lines, respectivly).

Each hydrological event in our dataset was categorized visually (by two or more domain experts) into one of the six hysteresis classes (Figure 3) of Williams (1989). Class I represents linear C-Q relationships that show little hysteretic behavior, whereas Class II and Class III represent clockwise and counter-clockwise hysteretic behaviors, respectively. A C-Q plot exhibiting a linear relationship followed by a clockwise loop is indicative of Class IV behavior. These patterns could reasonably be considered a special case of Class II (clockwise hysteresis); and rarely are studied as a separate hysteresis category (Malutta et al., 2020). The figure-eight loops are represented as Class V. Events that do not fall into any of these five classes are placed into a class labeled "Complex".

3.3 Multivariate Time Series Clustering

Clustering of the multivariate time series data at the storm event scale was a first step in exploring linkages between storm event responses (i.e., C-Q dynamics) and watershed processes. To this end, a number of clustering methods were investigated. Paparrizos and Gravano (2017) conducted extensive benchmark tests using four clustering algorithms (partitional, hierarchical, spectral, and density-based) and three distance measures – Euclidean distance, dynamic time warping of Sakoe and Chiba (1978), and shape-based (Paparrizos and Gravano, 2016). All of the datasets (85 in total) available in the University of California at Riverside (UCR) time series archive (Dau et al., 2018) at the time of their publication were used in the benchmark; they identified K-medoids with dynamic time warping (DTW) (discussed in Section 3.3.1 and Section 3.3.2, respectively) as having achieved the highest adjusted Rand index across the greatest number of datasets. Leveraging their work, we conducted additional benchmark tests using the four algorithms on their short list — TADPole (Begum et al., 2015), K-shape (Paparrizos and Gravano, 2016), K-medoids with DTW, and K-medoids with Euclidean. Using all datasets (currently 128 in total) available in the UCR time series archive (Dau et al., 2018), we also found that K-medoids with DTW achieved the highest adjusted Rand index across the greatest number of datasets. All of the event time series data in UCR archive were pre-processed as outlined in Section 3.1 to avoid unexpected consequences that might result from treating benchmark data differently from our hydrological event dataset.

3.3.1 K-medoids Clustering Algorithm

K-medoids is a variant of the popular K-means (Wu et al., 2007), in which the cluster centroids are observation points (called "medoids") as opposed to coordinates as in K-means. These medoids are mapped from a multivariate time series of length n (i.e., $t_1, t_2, ..., t_n$) to vectors of the multiple variables (i.e., $V1, V2, ..., Vm$) at each time step t_i. Like K-means, the K-medoids algorithm is iterative (Supporting Information Algorithm S1) where the initial K medoids are selected randomly. The algorithm has two phases: Phase 1 assigns observation points to clusters (Line 3); and Phase 2 calculates new medoids for each cluster (Line 4). In Phase 1, the distance between all observation points and each of the medoids is calculated, and each observation point is assigned to the closest medoid. In Phase 2, a new medoid is selected from each cluster by finding the observation point that minimizes the sum of squared distances (i.e., sum of squared errors) to all other observation points in that cluster. These two phases are repeated for a given number of iterations or until there is no change in the medoid selection. Algorithm S1 in Supporting Information was implemented in Python (version 3.6.1); the source codes may be found at GitHub (Javed, 2019b).

For a given dataset, the optimal number of clusters may vary depending on the research question/objective. In this study, the elbow method guided the selection of the "optimal" number of clusters. This method consists of plotting the sum of squared errors (SSEs) against an increasing number of K clusters. An optimal value for K is selected (visually) as the value for which further increases in K result in diminishing reduction in SSE, thus creating the onset of the plateau.

3.3.2 Dynamic Time Warping

The K-medoids clustering algorithm used a variant of dynamic time warping (DTW) to calculate the distance between two multivariate times series. Originally introduced for speech recognition (Sakoe and Chiba, 1978), DTW is arguably the most popular distance measure for time series clustering, and is particularly appealing for sensor data generated during hydrological events because of (i) the challenges associated with defining the beginning and end of an event (i.e., the ambiguity inherent in event delineation), and (ii) the noise present in the sensor data (e.g., variability in readings due to sensor interference from debris, maintenance activities, and temporary fouling.)

Figures 4a and 4b illustrate how distance between two time series ($T1$ in red and $T2$ in blue) is calculated using the more common Euclidean distance compared with DTW. While Euclidean distance uses a one-to-one alignment, DTW employs a one-to-many alignment that enables a warping of the time dimension to minimize the distance between the two time series. As such, DTW can optimize alignment, both global alignment (by shifting the entire time series left or right) and local alignment (by stretching or squeezing parts of time series). Paparrizos and Gravano (2016) showed that the best accuracy (as measured by the Rand index) is obtained when DTW is constrained to a limited window size. Multiple window size constraints ranging from 0% to 100% were tested to cluster our Mad River dataset. Based on a preliminary qualitative analysis of event visualizations, a window size constraint of 10% was selected for our analysis. Constraining the window size to 10% of the observation data is usually considered adequate for real applications (Ratanamahatana and Keogh, 2004); and it accommodates minor differences in timing between similar hydrological events, as is often the case when delineating the end of an event proves challenging.

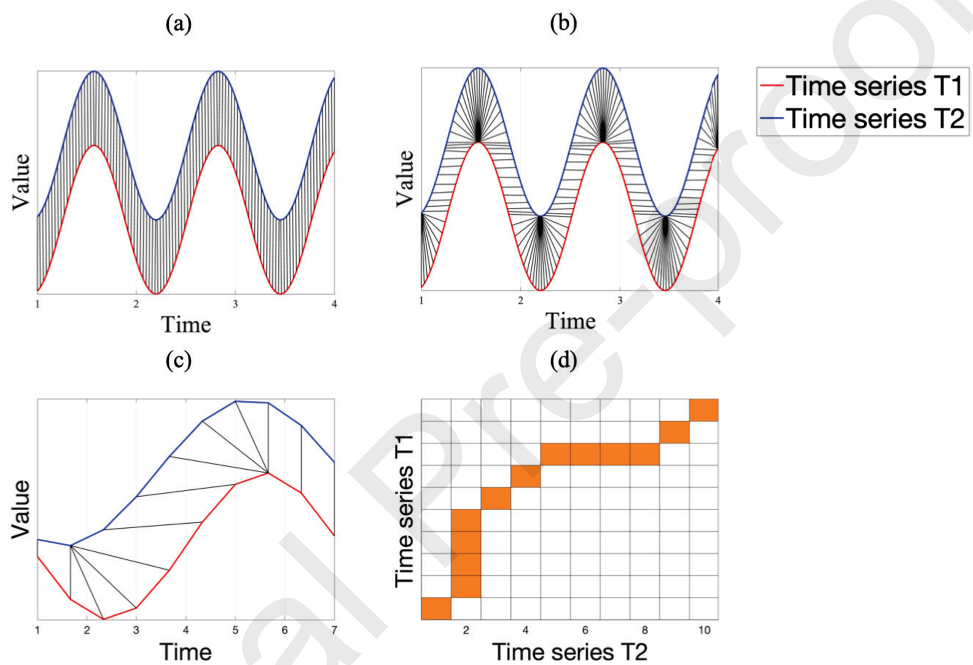


Figure 4: The top row illustrates the alignment between two times series for calculating distance in (a) Euclidean (one-to-one) and (b) dynamic time warping (one-to-many); Bottom row illustrates an optimal (c) alignment of each point in time series T1 and time series T2 (shown with black lines) and (d) warping path, i.e., optimal alignment of time series $T1$ (red) and $T2$ (blue), where each matrix cell (i, j) is the distance between ith element of $T1$ and jth element of $T2$; the DTW distance is the sum of the distances along the optimal path shown in orange.

Aligning two time series, $T1$ of length a and $T2$ of length b, using DTW involves creating an $a \times b$ matrix,

D, where the element $D[i,j]$ is the square of the Euclidean distance, $d(t1_i, t2_j)^2$, $d(\cdot,\cdot)$ is the Euclidean distance, $t1_i$ is the ith point of $T1$, and $t2_j$ is the jth point of $T2$. A warping path P is defined as the sequence of matrix elements that are mapped between $T1$ and $T2$ (see Figures 4c and 4d). This warping path must satisfy the following three conditions:

1. Every point from $T1$ must be aligned with one or more points from $T2$, and vice versa.
2. The first and last points of $T1$ and $T2$ must align, meaning the warping path must start and finish at diagonally opposite corner cells of the optimal warping matrix.
3. No cross-alignment is allowed, that is, the path must increase monotonically within the matrix.

For all paths that satisfy the three conditions above, DTW finds a path that minimizes the distance calculated as in Equation 1 (Shokoohi-Yekta and Keogh, 2015):

$$\mathrm{DTW}(T1, T2) = \min_{P \text{ mapping between } T1 \text{ and } T2} \sqrt{\sum_{(i,j)\in P} D[i,j]}, \tag{1}$$

Algorithm S2 in Supporting Information outlines the procedure for calculating this minimum distance using dynamic programming method (Bellman, 1957).

The environmental sensor data in this proof-of-concept are bivariate, representing water quality concentration and stream discharge time series. There are two DTW variants – DTW-independent (DTW-I) and DTW-dependent (DTW-D). In DTW-I, the distance between $T1$ and $T2$ is the sum of distances calculated separately for each variable (by invoking the DTW algorithm for each variable). Whereas in DTW-D, $T1$ and $T2$ are handled as *multivariate* time series; and the DTW algorithm is invoked only once. Because of the strong dependency between discharge and concentration in this work, DTW-D is used. The source code, implemented in Python (version 3.6.1), may be found at GitHub (Javed, 2019a).

3.4 Generating Synthetic Hydrograph and Concentration-graph Data

Synthetic multivariate times series "event data" were generated using eight conceptual hydrographs and two conceptual concentration graphs (Figure 5), and then combined to produce a set of heterogenous, albeit

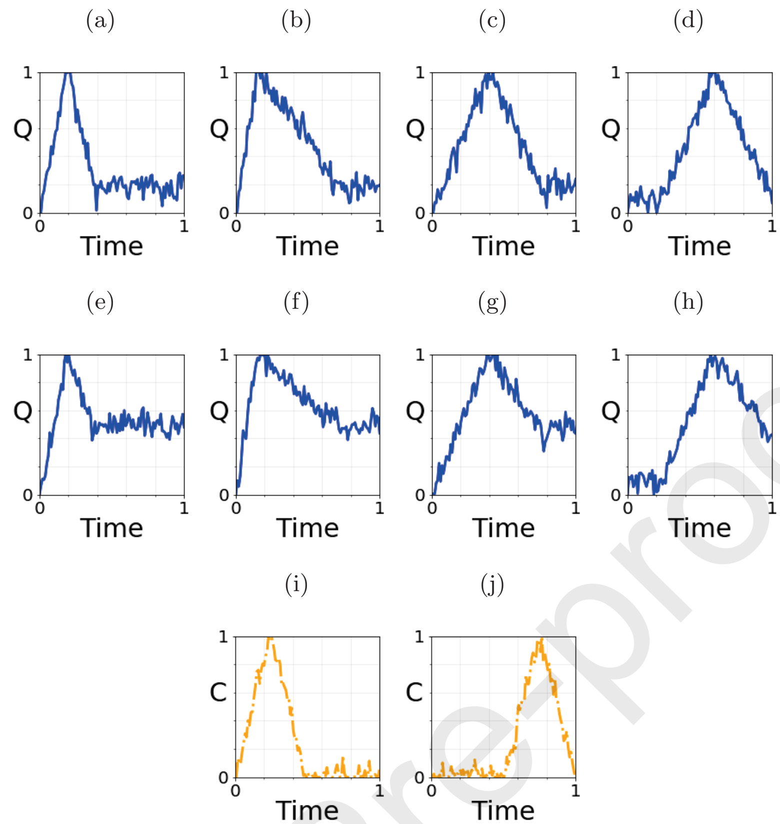


Figure 5: Example synthetic hydrographs and concentration graphs generated from eight conceptual hydrograph types: (a) flashy, early peak – return to baseline flow, (b) early peak – slow return to baseline flow, (c) mid-peak – return to baseline flow, (d) delayed rise to peak – return to baseline flow, (e) flashy, early peak – incomplete return to baseline flow, (f) early peak – slower incomplete return to baseline flow, (g) mid-peak – incomplete return to baseline flow, and (h) delayed rise to peak – incomplete return to baseline flow, and two conceptual concentration graphs: (i) early peak and (j) late peak.

simplified, hydrographs and sedigraphs (concentration graphs). A stochastic generator was designed to produce synthetic data with sensor noise. Random samples were drawn from a normal (Gaussian) distribution with a mean of 0.00 and standard deviation of 0.05 and added to the discharge and concentration values at each time step in order to simulate noise. When combining each of the eight synthetic hydrograps with the two concentration-graphs, sixteen synthetic storm event types can be produced. These combined event types can be labeled and used as "ground truth" events to help assess and validate the methodology.

Five control parameters, ranging from 0 to 1, were used to generate the synthetic graphs: time-to-peak, duration-of-peak, delay, recess, and initial baseline conditions. Time-to-peak controls the timing for the concentration/discharge values to reach the peak (normalized value of 1); duration-of-peak controls the duration of flow above baseline conditions; delay controls the time at which the value (either discharge or concentration) begins to rise in magnitude above the baseline conditions; recess controls the degree to which event concentration/discharge values return to the baseline conditions; and initial baseline controls the minimum value of the flow over an event. Parameter values for generating each type of synthetic graph (hydrograph and concentration-graph) were determined qualitatively based on re-production of simplified yet realistic approximation of typical hydrographs and sedigraphs observed in our study watershed (Supporting Information Table S2).

3.5 Measures for Assessing Clustering Performance

We used the *Hopkins Statistic* to measure the clustering tendency of our three datasets (i.e., the synthetic dataset, the Mad River dataset and the expanded regional Vermont dataset). The statistic value ranges from 0 to 1, where 1 indicates a high tendency to cluster and 0 indicates uniformly distributed data (Banerjee and Dave, 2004). Additionally, transformed variables (those representing the 24 storm event metrics of Table 2) were examined post-clustering to see whether these event metrics had 1) any association with clusters or 2) statistical power to differentiate between clusters using One-way Analysis of Variance (ANOVA) followed by Tukey Honest Significant Differences (HSD) tests between individual group means. For those variables (or their transformations) that were not normally distributed, nonparametric methods were applied (Kruskal-Wallis). Lastly, *Z-score* values were calculated for each of the 24 storm event metrics of Table 2 to identify feature importance associated with cluster differences. The Z-score represents the distance of an individual storm metric from the population mean (measured in terms of standard-deviation).

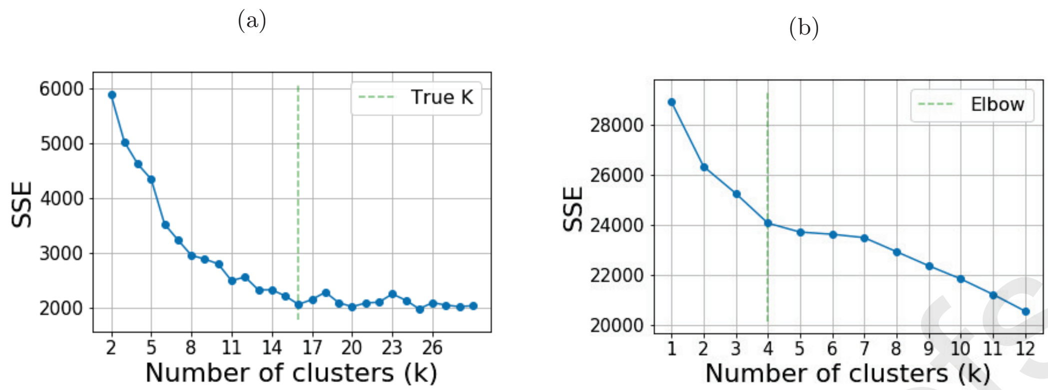


Figure 6: Sum of squared errors (SSE) for different number of clusters from (a) the synthetic storm event dataset (elbow point at K=16) and (b) the Mad River storm event dataset (elbow point at K=4).

4 Results

4.1 Using Synthetic Data to Validate Methodologies

To help validate the METS clustering approach, we generated 800 synthetic storm events, equally distributed among the sixteen possible combinations (see Section 3.4). As one might expect, the synthetic data had a high clustering tendency (Hopkins statistic of 1.00); and the optimal number of clusters, determined using elbow method as $K = 16$ (see Figure 6a), matched the intended synthetic design (16 event types). Examples of synthetic events from each of the 16 event classes are shown in Figure 7. Despite the presence of stochastically generated noise, the synthetic dataset clustered with 100% accuracy using K-medoids with DTW (i.e., clusters were identical to the ground truth).

4.2 Application of METS to the Mad River Dataset

In applying the METS clustering to the 603 Mad River storm events, we identified $K = 4$ event clusters with distinct SSC and Q responses (see the plateau in the elbow plot of Figure 6b). Approximately one third of the events (n = 234) fell into cluster 1, with each of the three remaining clusters having between 116 and 128 events (see Figure 8). Unlike the synthetic dataset, the optimal number of clusters for the Mad

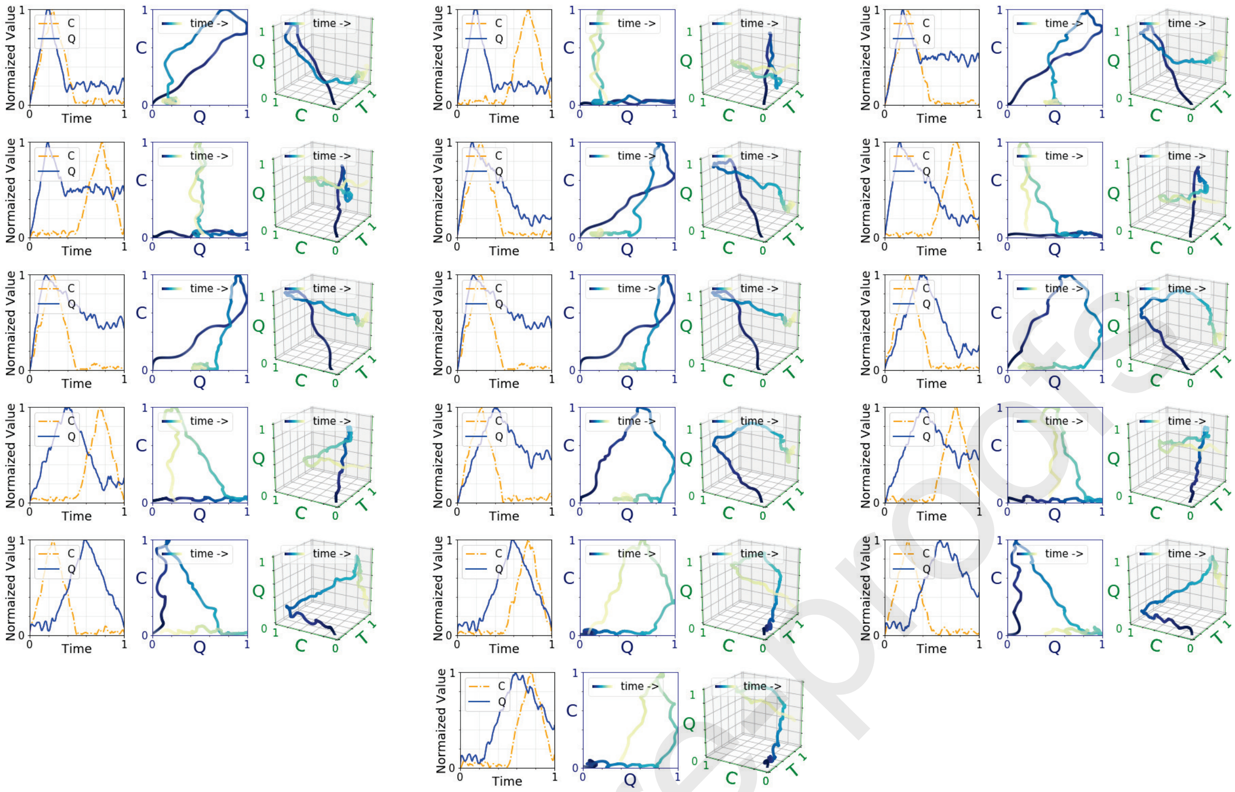
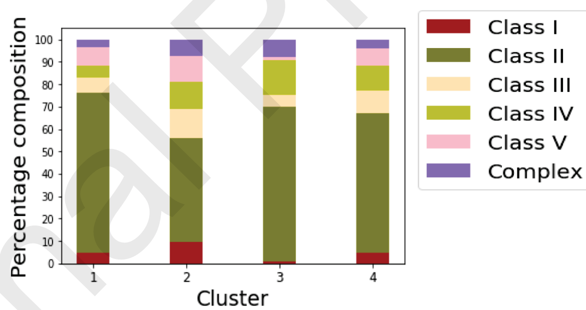


Figure 7: Example events in each of the 16 event classes in the synthetic dataset.



Cluster	Class I	Class II	Class III	Class IV	Class V	Complex	Total
1	11	167	16	12	20	8	234
2	12	58	16	15	15	9	125
3	1	80	6	18	2	9	116
4	6	80	13	14	10	5	128
Total	30	385	51	59	47	31	603

Figure 8: Distribution of hysteresis loop classes over METS clusters.

River dataset, any real dataset for that matter, will never be known with any degree of certainty. However, these data have a Hopkins test statistic of 0.96 indicating they are highly clusterable. We first explored whether a relationship existed between the four METS clusters and the six-class hysteresis scheme presented in Section 3.2. We found little association between the two as the confusion matrix and cluster distribution of Figure 8 show the six classes to be fairly evenly distributed across the four METS clusters.

4.2.1 Qualitative interpretation of METS clusters using event visualizations

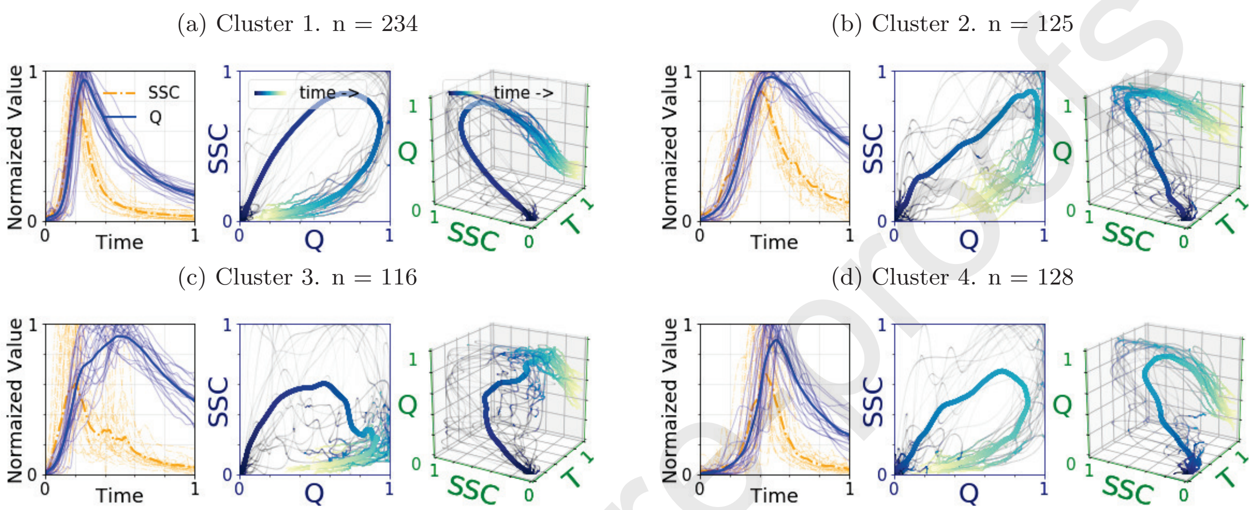


Figure 9: Mad River storm events closest to the centroid of each of the $K = 4$ clusters, superimposed on a single graph with the mean value plotted as a solid line — (a) cluster 1 events have a broad clockwise hysteresis pattern featuring an early and relatively brief duration of high SSC, (b) cluster 2 events have a narrow clockwise hysteresis loop and broad sedigraphs and hydrographs with streamflows that do not fully return to baseline levels, (c) cluster 3 events have flashier and sometimes multi-peaked sedigraphs that are shorter in duration, and (d) cluster 4 have a delayed rise of hydrograph and sedigraph, and typically more aligned.

Finding little relationship between the METS clustering and the hysteresis classification, we further investigated the characteristics associated with combined hydrograph and sedigraph trajectories of the METS clusters using multiple visualization approaches. To visualize overall trends, we superimposed 20 storm events closest to the centroid of each of the four METS clusters onto single plots (Figure 9); mean values are plotted as solid lines. Additionally, examples of the event times series, C-Q hysteresis plots, and 3-dimensional C-Q-T plots for each cluster are provided in Figure 10. In general, the METS cluster 1 events (Figure 9a and Figure 10a) have broad clockwise hysteresis patterns with an early, and relatively brief duration of high SSC. The hydrographs are flashy, rise quickly and return nearly to baseline flows. Cluster 2 events typically have a more narrow hysteresis loop compared to cluster 1 and broad (less flashy) sedigraphs and hydrographs with streamflows that do not fully return to the baseline levels (Figure 9b and Figure 10b). Cluster 3 events are similar to cluster 2, but exhibit flashier and sometimes multi-peaked sedigraphs that are shorter in duration (Figure 9c and Figure 10c). Multi-peaked events sometimes exhibit compound behavior including, for example, portions of clockwise hysteresis loops and no hysteretic behavior (linear relationships). Cluster 4 events typically have a delay in the rise of the hydrograph and sedigraph, and typically more aligned (Figure 9d and Figure 10d). In contrast to cluster 2 and 3 events, the hydrographs of cluster 4 also tend to return to near baseline levels.

4.2.2 Statistical Analysis of METS clusters

Table 3: Result of post-hoc Tukey HSD test ($\alpha = 0.05$) for all pairwise comparisons of hydrograph/sedigraph related storm event metrics. Within each metric, if two classes/clusters do not share the same letter, the metric means are significantly different.Shaded columns are highlighted to show examples of metrics distinguished well by METS, but not by hysteresis classes (light shading) and metrics discriminated well by hysteresis classes (dark shading).

Hydrograph/Sedigraph Characteristics									
Metric	T_Q	T_{SSC}	T_{QSSC}	Q_{Recess}	SSC_{Recess}	D_Q	FI	SSC_{Peak}	HI
METS clusters									
cluster 1	a	a	a	a	a	a	a	a	a
cluster 2	b	b	a	b	b	a b	b	b	b
cluster 3	b	c	b	c	a	a	b	b c	b
cluster 4	c	b	a	d	c	b	b	a c	b
Hysteresis classes									
Class I	a b	a b	a	a b	a b	a b	a b	a	a
Class II	a	a	b	a	a	a	a b	a	b
Class III	a	a	c	a	b	a b	a b	a	c
Class IV	a b	a b	a b	b	a	a b	a	a	d
Class V	a	a	a	a	a	a b	b	a	a
Complex	b	b	a b	a b	a	b	a b	a	a

Of the 24 storm event metrics in Table 2, 19 metrics had significantly different mean values for at least one of the METS clusters. The reader should bear in mind that these event metrics were not used as input to either the METS clustering algorithm or the hysteresis classification scheme. Both the METS clusters and hysteresis classes have event metrics with good discriminatory power; but there was little overlap for a given metric. For instance, two of the metrics shaded in Table 3 (e.g., SSC_{Peak} and the difference in discharge values at the beginning and end of an event (Q_{Recess})) show an ability to discriminate between the clusters generated by METS, but little statistical power to discriminate between the six classes of the hysteresis classification method. In contrast, both the hysteresis index (HI) and time between peak SSC and peak flow

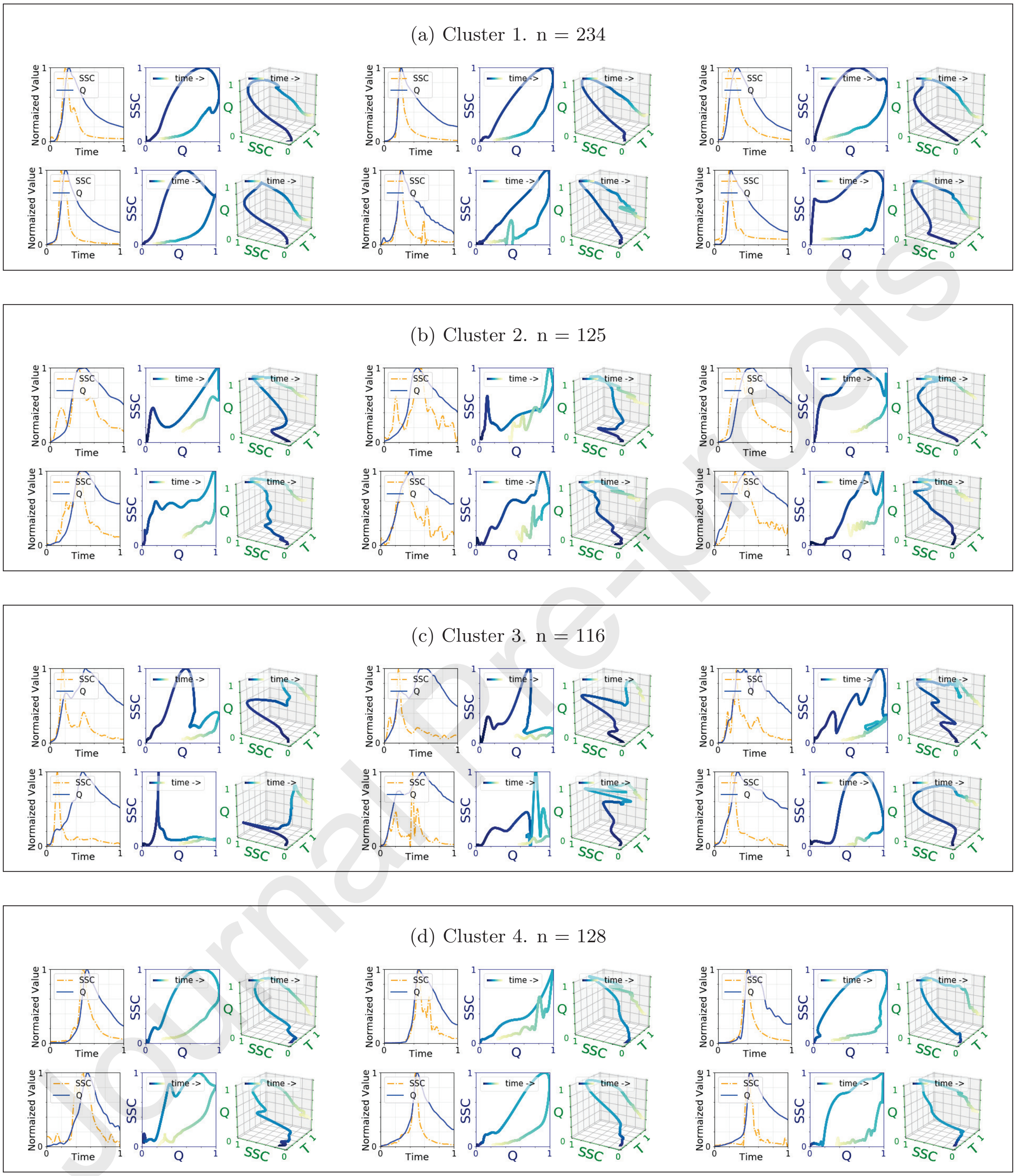


Figure 10: Six storm events closest to the centroid of the four Mad River dataset METS clusters ($K = 4$, $N = 603$) — (a) cluster 1 events have a broad clockwise hysteresis pattern featuring an early and relatively brief duration of high SSC, (b) cluster 2 events have a narrow clockwise hysteresis loop and broad sedigraphs and hydrographs with streamflows that do not fully return to baseline levels, (c) cluster 3 events have flashier and sometimes multi-peaked sedigraphs that are shorter in duration, and (d) cluster 4 have a delayed rise of hydrograph and sedigraph, and typically more aligned.

(T_{QSSC}) show power to discriminate between the hysteresis classes, but not the MET clusters (Table 3). Similar differences in discriminatory power were observed in metrics related to antecedent conditions, rainfall characteristics, and streamflow/sediment characteristics (Supporting Information Table S3 to Table S5).

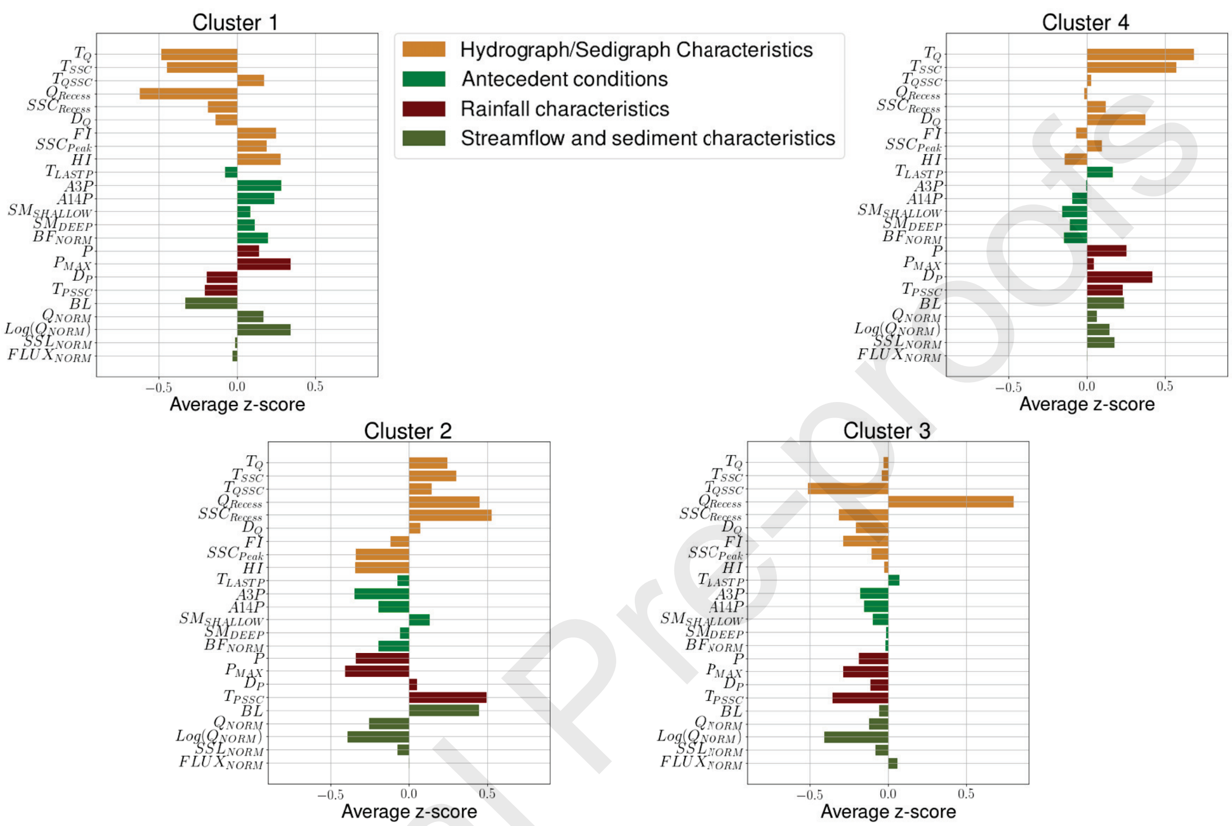


Figure 11: Typical hydrometeorological characteristics of METS clusters as represented by storm event Z-score metrics for each of the four clusters.

Next, we explored the hydrometeorotological factors associated with the four METS clusters using event metric Z-score values. Again, these event metrics were not used as input to the clustering algorithm, but as a means to study linkages between these characteristics and the resulting clusters. The storm events of cluster 1 have greater amounts of precipitation (positive Z-score for P and P_{Max}) and wetter antecedent conditions exhibited by higher mean BF_{Norm}, SM_{Deep},$SM_{Shallow}$, $A3P$ and $A14P$. In general, these factors are associated with higher stream discharge as confirmed by the positive Z-score for Log (Q_{Norm}), Q_{Norm}, and FI (flood intensity) as well as higher peak SSC values. Other notable characteristics include hydrographs that return to baseline flow (negative Z-score for Q_{Recess}), and a rapid rise in the sedigraph and hydrograph (negative Z-score for T_{SSC} and T_Q) and positive Z-score for HI, which translate to a 2D hysteresis that is dominated by a broad clockwise pattern (observed in Figure 9a and Figure 10a).

Cluster 2 is associated with smaller precipitation events (negative Z-score for P and P_{Max}) and drier antecedent conditions (negative BF_{Norm}, SM_{Deep}, $A3P$ and $A14P$ Z-scores), both resulting in lower stream discharge (negative Log (Q_{Norm}), Q_{Norm}, and FI Z-scores). These events also have positive Q_{Recess} and SSC_{Recess} Z-score values. These two metrics were designed to capture whether streamflow and SSC return to baseline levels; positive scores are associated with events that do not return to base levels (Figure 9b and Figure 10b). Additional characteristics include lower peak SSC concentrations and negative Z-scores for BL (indicative of watersheds that respond more slowly to a rainfall event), and a longer duration between the peak SSC and center of mass for rainfall (positive Z-score for T_{PSSC}). The latter translates to hysteresis patterns with more narrow loop, which is confirmed visually (Figure 9b and Figure 10b), and by the negative Z-score for hysteresis index.

Cluster 3 events have a rapid rise in both streamflow and SSC (Figure 9c and Figure 10c) and are associated with a positive Z-scores for Q_{Recess} and negative for SSC_{Recess}, which is indicative of sedigraphs that return to base levels and hydrographs that do not. The sedigraph is also often characterized by multiple peaks; and in general, there is a short duration between the peak SSC and the center of mass for rainfall (negative Z-score for T_{PSSC}) as well as between the peak SSC and peak discharge (negative T_{QSSC}). In addition, these events have lower precipitation (negative Z-scores for P and P_{Max}) and stream discharge (negative Log (Q_{Norm}), Q_{Norm}, and FI), as well as Z-scores that approach zero for BF_{Norm}, SM_{Deep}, $SM_{Shallow}$, $A3P$ and $A14P$, which indicate average antecedent conditions.

Lastly, cluster 4 events are associated with higher precipitation (positive Z-score for P) that are longer in duration (positive Z-score for D_P); however, these events have less intense rainfall (near zero Z-score for P_{Max}), and are associated with average to fairly dry antecedent conditions (i.e., slightly negative Z-score values for BF_{Norm}, SM_{Deep}, $SM_{Shallow}$, $A3P$ and $A14P$), all of which results in near average streamflows (near zero Z-score for Log (Q_{Norm}), Q_{Norm}, and FI). Other event characteristics include a long time to peak SSC and Q (positive Z-score for T_{SSC} and T_Q) and larger amounts of sediment transport during events (positive SSL_{Norm}).

4.3 Effects of Additional Watersheds on METS Clustering

The number and type of event clusters/classes are dependent on geographic range of study. In re-running the METS analysis on the expanded regional Vermont dataset, the number of clusters increased from $K = 4$ to $K = 9$ (Supporting Information Figure S2). This is not surprising given the differences, particularly in topography and land use, associated with the added watersheds. Hungerford Brook, for instance, is a low gradient agricultural basin, while Allen Brook drains a highly developed suburban area (Supporting Information Table S1). The METS results show the expanded dataset cluster 5 to have a substantially large number (54%) of counter-clockwise hysteresis loops, which correspond to events where the sedigraph peaks after the hydrograph (hysteresis Class III), and no events that are clockwise (hysteresis Class II or Class IV) (Supporting Information Figure 12 and Table S6).

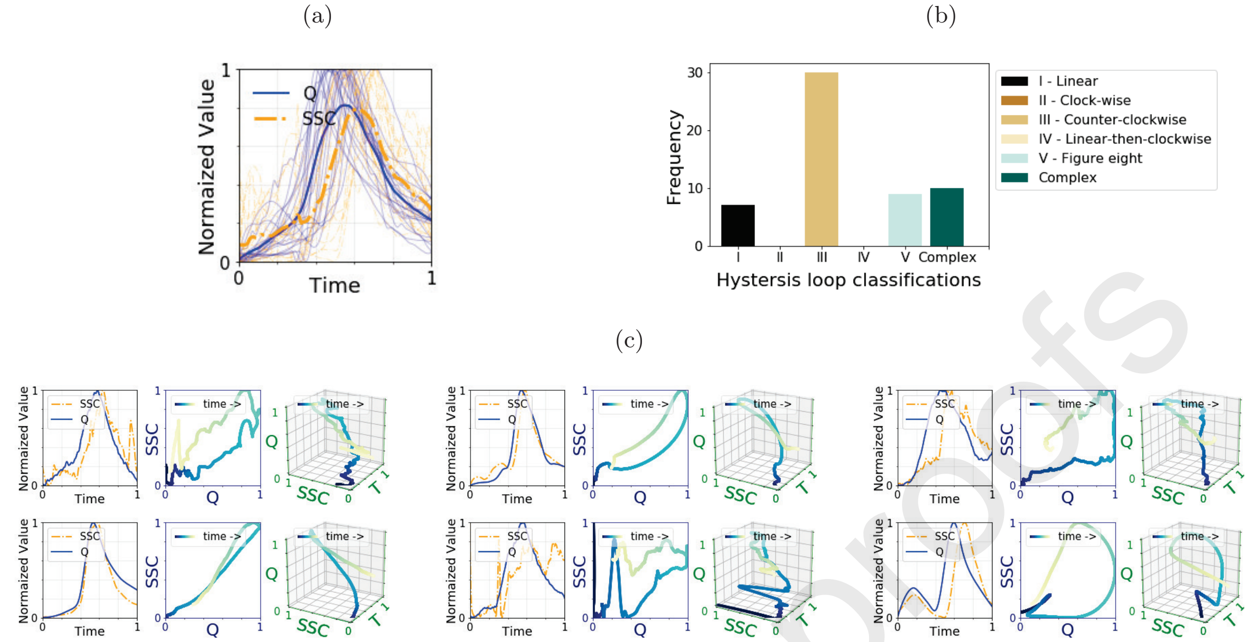


Figure 12: Storm events closest to the centroid of the cluster 5 dominated by counter clockwise hysteresis type events (when $K = 9$) in the expanded regional Vermont dataset, discovered by including more watersheds: (a) all 56 events in cluster 5 superimposed, with the mean plotted as a solid line, (b) distribution of cluster by hysteresis loop classification, and (c) six events closest to the centroid of the cluster ($n = 56$).

5 Discussion

We present a new clustering approach within the broader discipline of event-based studies — one that leverages the temporal information in two or more time series for the purpose of grouping or identifying similar events — in this manuscript, a *hydrological* event comprising hydrograph and sedigraph data modeled as three-dimensional C-Q-T trajectories. This contrasts with current hydrological event approaches that either collapse the time dimension (e.g., 2D hysteresis pattern analysis of Lloyd et al. (2016b)) or focus on the response of a single variable such as the DTW clustering approach of Dupas et al. (2015); the latter re-scales events using a single (ideal) hydrograph and then clusters the concentration response. While these approaches are important to a variety of research applications, these 2-D hysteresis methodologies lose the temporal information, while the latter requires a rescaling of the C-Q variables. The multivariate version of DTW-D used in the METS clustering of this manuscript is designed to extract relationships between the time series of two or more variables, resulting in a dataset partitioning that is dissimilar and complementary to existing hysteresis methods.

5.1 Effects of Regional Scale on METS Clustering.

Our motivations for limiting the primary analysis to the Mad River watershed were two-fold. First, meteorological data were not available for the additional watersheds; and secondly, we wanted, at least initially, to control for certain watershed characteristics such as topography and land use (e.g., the Mad River has primarily two land use types - forest and agriculture). In this single watershed study, we identified four predominant clusters for hydrological events occurring between the period from 2013 and 2016, with one cluster type occurring most frequently (38%), and 64% of the events categorized as clockwise patterns. This relatively small number of event types (i.e., four clusters) might be expected, given the uniformity of watershed characteristics across the six Mad River monitoring sites; as this is similar in number to other event analyses from single study areas. Bende-Michl et al. (2013) identified 3-4 cluster in a study on nutrient dynamics; Mather and Johnson (2015) identified 5-7 clusters when analyzing C-Q loops; and 3 nutrient-event response types were identified in the work of Dupas et al. (2015). In general, there is a great deal of interest and merit in tracking the change in both the number and type of event responses within a single study area, particularly for example, when monitoring in-stream changes prior to and after restoration efforts. However, other monitoring applications may require tracking changes across watersheds at larger geographical scale; and one might expect the number of clusters (event types) to increase with the geographic range of study as demonstrated in Section 4.3.

Regardless of regional scale, we found the METS clustering to be heavily influenced by the degree to which both of the time series (SSC and Q) return (or not) to base levels at the end of the event. This was evidenced both visually (Figure 10) and by the significance of the SSC_{Recess} and Q_{Recess} metrics (Table 3 and Figure 11). From a hydrological perspective, the rate and degree of recession (return to baseline flow and background concentration levels) are important indicators of soil moisture, groundwater elevations, and the resulting hydrological flowpaths. Classification schemes based on the shape and direction of hysteresis do not necessarily capture this "return to baseline conditions" behavior because the overall C-Q patterns are primarily driven by the middle portion of the hydrograph-sedigraph (i.e. largest offset between C-Q) rather than differences between the times series at the start or end of the event. The ability of the METS clustering to capture this return-to-baseline conditions phenomena, in addition to other metrics, holds promise for many applications (e.g., model validation) used in forecasting floods, water quality monitoring, watershed similarity studies, and detecting change in watershed functions.

5.2 Leveraging Methodological Strengths to Group Events

The post-cluster analysis performed on event metrics (hydrological and meteorological metrics in Table 2) was an attempt to explore which factors (i.e., characteristics associated with the event time series) might be driving the METS clustering, bearing in mind that these metrics were not used as inputs to the clustering analysis itself. Prior event-based hydro-meteorological studies have successfully used this type of post-statistical analysis to tease out factors important in discriminating between (or correlated with) event groupings. Examples include the classifying of event hysteresis patterns to study erosional processes (Seeger et al., 2004; Nadal-Romero et al., 2008; Sherriff et al., 2016; Hamshaw et al., 2018).

Here, we highlight some key results from our post-cluster statistical analysis, particularly the event metric with statistically significant differences across the METS clustering and/or hysteresis classification. First, while the event hysteresis index (HI) was identified, not surprisingly, as important for differentiating between the hysteresis class types (see Table 3 in Supporting Information), the temporal hydrograph and sedigraph metrics (e.g., time to peaks – T_Q, and T_{SSC}), as well as the degree to which both time series return to baseline conditions (Q_{Recess} and SSC_{Recess}) were not identified as important drivers. In contrast, these four metrics as well as the Peak SSC (SSC_{Peak}), duration of stormflow (D_Q) and antecedent precipitation metrics (Section 4.2.2) were identified as important for differentiating between the METS-based clusters (Table 3 and Supporting Information Table S3).

5.3 Using Methods in Tandem to Leverage Strengths

(a) (b)

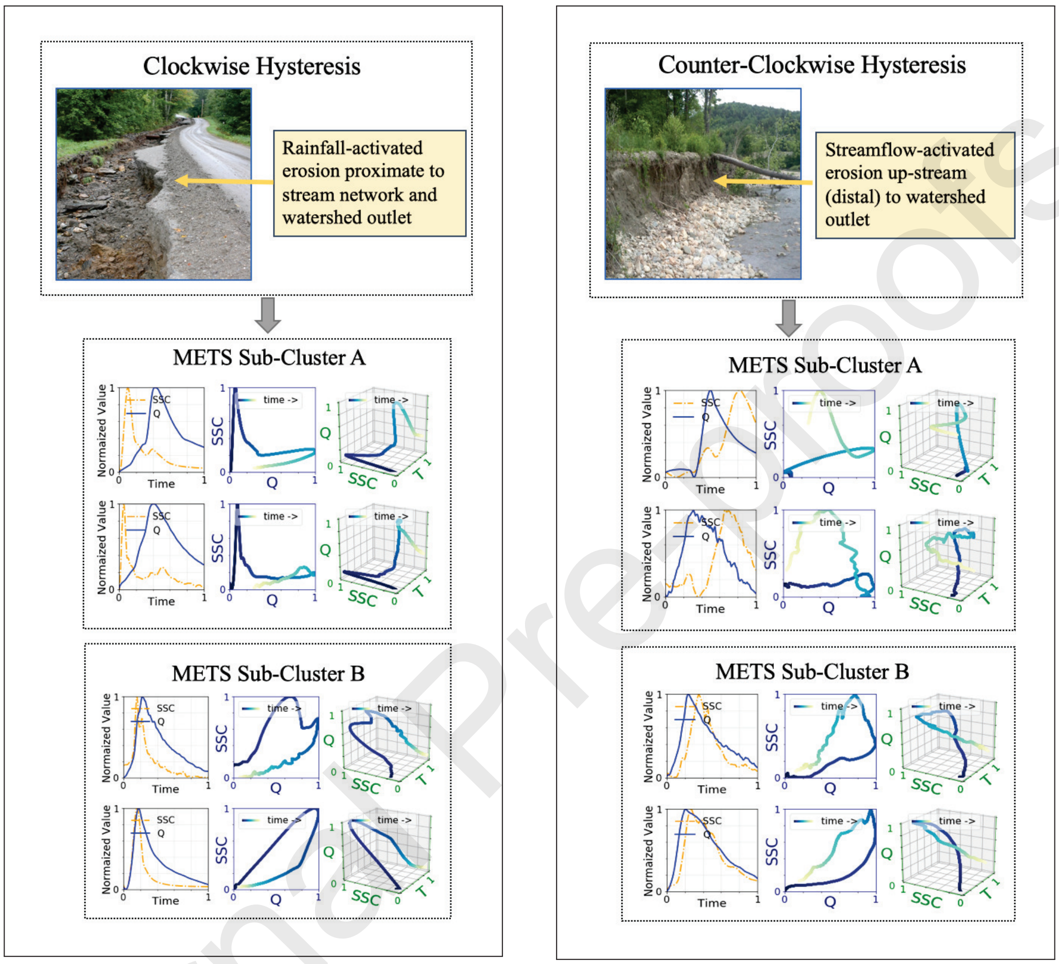


Figure 13: Application of METS after pre-classifying events based on hysteresis directions of (a) clockwise hysteresis and (b) counter clockwise hysteresis that can correspond to general proximity and timing of erosion source activation. METS clustering further partitions these hysteresis classes into sub-clusters (visualized as two example events) distinguishable by different hydrograph and sedigraph characteristics. Photos from observed, active erosion sources within the Mad River watershed.

Each of the clustering and classification approaches have unique strengths and weaknesses; and the post-statistical analyses (e.g., Tukey HSD test and Z-scores of Section 4.2.2) provide some guidance on method selection that best aligns with manager or stakeholder goals. However, using more than one method in tandem may help to leverage methodological strengths. For example, in event-based suspended sediment studies — those aimed at identifying the proximity of riverine erosion sources, a two-phased approach may add value. Let's consider our expanded dataset in which more than two thirds of the events have clockwise hysteresis patterns. A first phase might use hysteresis classification to prioritize the clockwise versus counter-clockwise nature of the hysteresis patterns, as the direction embeds key process information. This Phase I classification could then be further partitioned into subgroups (via METS methodology) to help refine the understanding of watershed processes.

To highlight the potential of such an approach, we applied the 2-D hysteresis analysis and METS clustering in tandem using the expanded dataset of Section 4.3. In Phase I, hydrological events were classified (e.g., into clockwise and counter-clockwise groups) based on their hysteresis patterns; and in Phase II, the METS clustering was applied to each of the Phase I classes, respectively (Figure 13 and Supporting Information Figure S3 and Figure S4). Clockwise hysteresis patterns are typically indicative of erosion sources (e.g., gullies or rills) that are located very close to the monitoring site. Whereas the events in the counter-clockwise group are characterized by hydrographs that occur (and peak) prior to the accompanying sedigraphs. These are often indicative of more distal sediment sources (e.g., upstream streambank collapse). The METS sub-clusters shown in the lower half of Figure 13 (sub-clusters B), were differentiated by temporal information that was not fully captured by the Phase I hysteresis classification. Both sub-clusters are characterized by hydrographs and sedigraphs that return more completely (relative to sub-clusters A) to baseline levels. Whether used on its own or on a dataset that has been pre-classified or grouped by some other means, METS offers hydrological researchers a flexible and powerful approach for data-driven analysis of high-frequency water quality data; and the methodology may be easily adapted to different analysis objectives.

5.4 Challenges and Opportunities

The sparsity of hydrological events is an inherent data challenge that relies on data-driven or machine learning methods of analysis. Our study area, a typical humid and temperate watershed, experiences on average about 30 rainfall-runoff (i.e., storm) events a year. Other recent, prominent event-based studies (Wymore et al., 2019; Sherriff et al., 2016; Vaughan et al., 2017) are similarly constrained by event sizes ranging between 8 and 90 events per monitoring site. Albeit large from an environmental monitoring perspective, these relatively small sample sizes cause significant challenges for machine learning methods. The challenges are compounded when analyzing multivariate time series generated from in-situ sensors that must be kept online during extreme events and operating simultaneously. Currently, the hydrological informatics community is investing significantly in the integration and maintenance of data hubs that comprise multiple researchers across multiple organizations such as those of the Consortium of Universities for the Advancement of Hydrological Sciences, Inc. (CUAHSI, 2019). Despite the development of new machine learning methods to address data sparsity issues, another promising approach is to generate synthetic hydrological storm events as demonstrated in this work.

METS clustering operates on delineated events and is influenced by the degree to which both time series (SSC and Q) return (or not) to base levels at the end of the event. This highlights the importance of precise event delineation in METS clustering. In hydrology, many event-based studies rely on semi-automated and somewhat subjective methods to identify the start and end of an event, particularly when handling multipeak (consecutive) events (Wymore et al., 2019; Vaughan et al., 2017; Hamshaw et al., 2018; Sherriff et al., 2016; Gellis, 2013). Automation of event delineation is another area that can benefit from advances in machine learning methods, new data hubs, and access to synthetic, pre-delineated event data.

A key challenge with any clustering method is determining the optimal number, K, of categories (e.g., the correct number of storm event types). In this work, we select K based on the inflection point of an elbow plot. However, identifying the inflection point is often subjective. This is further complicated in hydrogeological applications, where the optimal number of categories is dependent on both the research objectives as well as the geographic location. In this proof-of-concept, we made no assumptions or preconceptions about the desired number of outcome categories. However, domain experts familiar with a particular region of study may have intuitive knowledge regarding the desired number of outcomes. Varying the number of clusters in METS is relatively straightforward and not computationally intensive; thus, researchers can easily evaluate the effect of cluster number – particularly when methods for evaluating "optimal" (e.g., the elbow method) are not definitive. Alternatively, one could replace the METS clustering algorithm with an alternative algorithm such as the density-based clustering algorithm of Ester et al. (1996), which does not require the number of clusters as an input.

The METS clustering approach is applicable to any water quality constituent or solute (e.g., nitrate, phosphorous and conductivity), which would be expected to demonstrate very different C-Q-T trajectories and resulting clusters compared to suspended sediment concentration response (Lloyd et al., 2016a; Zuecco et al., 2016). Additionally, the approach may be extended beyond a single parameter (e.g., SSC) to multiple parameters (e.g., SSC and nitrate) to explore/reveal any unknown interactions during storm events. Expansion to multiple parameters will bring interesting visualization and analysis challenges. One approach may be to visualize events as 3-D signal trajectories such as those we presented in this work.

6 Conclusion

The rapidly increasing volume and availability of high-frequency time series data offer considerable opportunity to analyze watershed systems at the storm event scale. In this work, we introduce the multivariate event time series (METS) approach for categorizing hydrological storm events into a limited number of clusters given data from multiple sensors deployed in the Mad River watershed in Vermont, USA. In order to validate the approach, we showed that stochastic generation of synthetic hydrographs and concentration graphs provided a simple and effective solution to over-coming the data sparsity challenge in training machine learning algorithms on environmental data. The approach is flexible enough to be used with any water quality constituents (e.g., nitrate, phosphorous and conductivity) alone or in combination. We highlight areas for further research to expand the application of event-based analysis. Additionally, we discuss how the METS clustering can be used in tandem with a traditional hysteresis based event classification scheme. Whether used on its own or in tandem with other partitioning methods, this method offers hydrological researchers a flexible and powerful approach for analyzing high-frequency water quality data; and opens up new possibilities for interpreting emergent event behavior in watersheds.

7 Acknowledgements

This project was supported by the Richard Barrett Foundation and Gund Institute for Environment through a Gund Barrett Fellowship. Additional support was provided by the Vermont EPSCoR BREE Project (NSF Award OIA-1556770). We thank Dr. Jae-Gil Lee, Associate Professor at the Korea Advanced Institute of Science and Technology (KAIST), for guiding the algorithm selection, and Dr. Patrick J. Clemins of Vermont EPSCoR, for providing support in using the EPSCoR Pascal high-performance computing server for the project.

References

Aguilera, R. and Melack, J. M. (2018). Concentration-discharge responses to storm events in coastal California watersheds. *Water Resources Research*, 54(1):407424.

Banerjee, A. and Dave, R. N. (2004). Validating clusters using the Hopkins statistic. In *Proceedings of the 2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No.04CH37542)*, volume 1, pages 149–153.

Begum, N., Ulanova, L., Wang, J., and Keogh, E. (2015). Accelerating dynamic time warping clustering with a novel admissible pruning strategy. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 49–58.

Bellman, R. (1957). *Dynamic Programming*. Dover Publications.

Bende-Michl, U., Verburg, K., and Cresswell, H. P. (2013). High-frequency nutrient monitoring to infer seasonal patterns in catchment source availability, mobilisation and delivery. *Environmental Monitoring and Assessment*, 185(11):9191–9219.

Burns, D. A., Pellerin, B. A., Miller, M. P., Capel, P. D., Tesoriero, A. J., and Duncan, J. M. (2019). Monitoring the riverine pulse: Applying high-frequency nitrate data to advance integrative understanding of biogeochemical and hydrological processes. *Wiley Interdisciplinary Reviews: Water*, page e1348.

Burt, T. P., Worrall, F., Howden, N. J. K., and Anderson, M. G. (2015). Shifts in discharge-concentration relationships as a small catchment recover from severe drought. *Hydrological Processes*, 29(4):498507.

Chen, L., Sun, C., Wang, G., Xie, H., and Shen, Z. (2017). Event-based nonpoint source pollution prediction in a scarce data catchment. *Journal of Hydrology*, 552:13–27.

CUAHSI (2019). Consortium of universities for the advancement of hydrologic science, inc. `https://www.cuahsi.org`.

Dau, H. A., Keogh, E., Kamgar, K., Yeh, C.-C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C. A., Yanping, Hu, B., Begum, N., Bagnall, A., Mueen, A., Batista, G., and Hexagon-ML (2018). The UCR time series classification archive. `https://www.cs.ucr.edu/~eamonn/time_series_data_2018/`.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines.* Oxford University Press, Inc.

Dupas, R., Tavenard, R., Fovet, O., Gilliet, N., Grimaldi, C., and Gascuel-Odoux, C. (2015). Identifying seasonal patterns of phosphorus storm dynamics with dynamic time warping. *Water Resources Research*, 51(11):8868–8882.

Ehret, U. and Zehe, E. (2011). Series distance - An intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events. *Hydrology and Earth System Sciences*, 15(3):877–896.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231.

Ewen, J. (2011). Hydrograph matching method for measuring model performance. *Journal of Hydrology*, 408(1):178 – 187.

Gellis, A. (2013). Factors influencing storm-generated suspended-sediment concentrations and loads in four basins of contrasting land use, humid-tropical Puerto Rico. *CATENA*, 104:39 – 57.

Hamshaw, S., M. Dewoolkar, M., W. Schroth, A., Wemple, B., and M. Rizzo, D. (2018). A new machine-learning approach for classifying hysteresis in suspended-sediment discharge relationships using high-frequency monitoring data. *Water Resources Research*, 54:4040–4058.

Javed, A. (2019a). Dynamic Time Warping. `https://github.com/ali-javed/dynamic-time-warping`.

Javed, A. (2019b). K-medoids for multivariate time series clustering. `https://github.com/ali-javed/Multivariate-Kmedoids`.

Keesstra, S. D., Davis, J., Masselink, R. H., Casal, J., Peeters, E. T. H. M., and Dijksma, R. (2019). Coupling hysteresis analysis with sediment and hydrological connectivity in three agricultural catchments in Navarre, Spain. *Journal of Soils and Sediments*, 19(3):15981612.

Latecki, L. J., Megalooikonomou, V., Qiang Wang, Lakaemper, R., Ratanamahatana, C. A., and Keogh, E. (2005). Partial elastic matching of time series. In *Proceedings of the 5th IEEE International Conference on Data Mining*, pages 4 pp.–.

Latecki, L. J., Megalooikonomou, V., Wang, Q., Lakaemper, R., Ratanamahatana, C. A., and Keogh, E. (2005). Elastic partial matching of time series. In *Knowledge Discovery in Databases*, pages 577–584.

Lloyd, C., Freer, J., Johnes, P., and Collins, A. (2016a). Using hysteresis analysis of high-resolution water quality monitoring data, including uncertainty, to infer controls on nutrient and sediment transfer in catchments. *Science of The Total Environment*, 543, Part A:388 – 404.

Lloyd, C. E. M., Freer, J. E., Johnes, P. J., and Collins, A. L. (2016b). Technical Note: Testing an improved index for analysing storm discharge–concentration hysteresis. *Hydrology and Earth System Sciences*, 20(2):625–632.

Malutta, S., Kobiyama, M., Chaffe, P. L. B., and Bonum, N. B. (2020). Hysteresis analysis to quantify and qualify the sediment dynamics: state of the art. *Water Science and Technology*. wst2020279.

Mather, A. L. and Johnson, R. L. (2014). Quantitative characterization of stream turbidity-discharge behavior using event loop shape modeling and power law parameter decorrelation. *Water Resources Research*, 50(10):7766–7779.

Mather, A. L. and Johnson, R. L. (2015). Event-based prediction of stream turbidity using a combined cluster analysis and classification tree approach. *Journal of Hydrology*, 530:751 – 761.

Minaudo, C., Dupas, R., Gascuel-Odoux, C., Fovet, O., Mellander, P.-E., Jordan, P., Shore, M., and Moatar, F. (2017). Nonlinear empirical modeling to estimate phosphorus exports using continuous records of turbidity and discharge. *Water Resources Research*, 53:7590–7606.

Nadal-Romero, E., Regs, D., and Latron, J. (2008). Relationships among rainfall, runoff, and suspended sediment in a small catchment with badlands. *CATENA*, 74(2):127 – 136.

Onderka, M., Krein, A., Wrede, S., Martinez-Carreras, N., and Hoffmann, L. (2012). Dynamics of storm-driven suspended sediments in a headwater catchment described by multivariable modeling. *Journal of Soils and Sediments*, 12(4):620–635.

Paparrizos, J. and Gravano, L. (2016). K-shape: Efficient and accurate clustering of time series. *SIGMOD Record*, 45(1):69–76.

Paparrizos, J. and Gravano, L. (2017). Fast and accurate time-series clustering. *ACM Transactions on Database Systems*, 42(2):8:1–8:49.

PRISM (2019). PRISM climate group. `http://prism.oregonstate.edu`. Last accessed on March 16, 2019.

Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., and Keogh, E. (2012). Searching and mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 262–270.

Ratanamahatana, C. A. and Keogh, E. (2004). Everything you know about Dynamic Time Warping is wrong. In *Proceedings of the 3rd Workshop on Mining Temporal and Sequential Data*. Citeseer.

Rose, L. A., Karwan, D. L., and Godsey, S. E. (2018). Concentration-discharge relationships describe solute and sediment mobilization, reaction, and transport at event and longer timescales. *Hydrological Processes*, 32(18):28292844.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

Savitzky, A. and Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36:1627–1639.

Seeger, M., Errea, M.-P., Beguera, S., Arnez, J., Mart, C., and Garca-Ruiz, J. (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. *Journal of Hydrology*, 288(3):299–311.

Sherriff, S. C., Rowan, J. S., Fenton, O., Jordan, P., Melland, A. R., Mellander, P.-E., and hUallachin, D. O. (2016). Storm event suspended sediment-discharge hysteresis and controls in agricultural watersheds: Implications for watershed scale sediment management. *Environmental Science & Technology*, 50(4):17691778.

Shokoohi-Yekta, M. and Keogh, E. J. (2015). On the non-trivial generalization of Dynamic Time Warping to the multi-dimensional case. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 289–297.

Stryker, J., Wemple, B., and Bomblies, A. (2017). Modeling sediment mobilization using a distributed hydrological model coupled with a bank stability model. *Water Resources Research*, 53(3):2051–2073.

Vaughan, M. C. H., Bowden, W. B., Shanley, J. B., Vermilyea, A., Sleeper, R., Gold, A. J., Pradhanang, S. M., Inamdar, S. P., Levia, D. F., Andres, A. S., and et al. (2017). High-frequency dissolved organic carbon and nitrate measurements reveal differences in storm hysteresis and loading in relation to land cover and seasonality: high-resolution doc and nitrate dynamics. *Water Resources Research*, 53:5345–5363.

Wemple, B. C., Clark, G. E., Ross, D. S., and Rizzo, D. M. (2017). Identifying the spatial pattern and importance of hydro-geomorphic drainage impairments on unpaved roads in the northeastern usa. *Earth Surface Processes and Landforms*, 42(11):1652–1665.

Wendi, D., Merz, B., and Marwan, N. (2019). Assessing hydrograph similarity and rare runoff dynamics by cross recurrence plots. *Water Resources Research*, 55(6):4704–4726.

Williams, G. P. (1989). Sediment concentration versus water discharge during single hydrologic events in rivers. *Journal of Hydrology*, 111(1):89–106.

Williams, M. R., Livingston, S. J., Penn, C. J., Smith, D. R., King, K. W., and Huang, C.-h. (2018). Controls of event-based nutrient transport within nested headwater agricultural watersheds of the western Lake Erie basin. *Journal of Hydrology*, 559:749–761.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., and Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37.

Wymore, A. S., Leon, M. C., Shanley, J. B., and McDowell, W. H. (2019). Hysteretic response of solutes and turbidity at the event scale across forested tropical montane watersheds. *Frontiers in Earth Science*, 7:126.

Zuecco, G., Penna, D., Borga, M., and van Meerveld, H. J. (2016). A versatile index to characterize hysteresis between hydrological variables at the runoff event timescale. *Hydrological Processes*, 30(9):1449–1466.

8 Supporting Information

This supporting information contains tables and figures to provide additional information on the following aspects of the study:

1. Table S1: Study watershed characteristics.
2. Figure S1: Matrix representation of multivariate time series.
3. Algorithm S1: K-medoids algorithm for hydrological event clustering.
4. Algorithm S2: Dynamic time warping algorithm for calculating the distance between two time series.
5. Table S2: Default parameter settings for synthetic hydrograph and concentration-graph generator.
6. Table S3: Result of post-hoc Tukey HSD test for all pairwise comparisons of antecedent conditions metrics.
7. Table S4: Result of post-hoc Tukey HSD test for all pairwise comparisons of rainfall characteristics metrics.
8. Table S5: Result of post-hoc Tukey HSD test for all pairwise comparisons of streamflow and sediment characteristics metrics.
9. Figure S2: SSE for varying number of clusters for Mad River dataset and Expanded dataset.
10. Table S6: Distribution of hysteresis loop classes over METS cluster 5 (when K= 9) in the expanded dataset(n= 56).
11. Figure S3: Three storm events closest to the centroid of the four extended dataset tandem clockwise hysteresis sub-clusters (K= 4, N= 496).
12. Figure S4: Three storm events closest to the centroid of the four extended dataset tandem counter clockwise hysteresis sub-clusters (K= 2, N= 90).

Table S1: Study watershed characteristics.

Characteristic	Freeman Brook	Folsom Brook	Mill Brook	High Bridge Brook	Shepard Brook	Mad River	Allen Brook	Hungerford Brook	Wade Brook
Area (km^2)	17.0	18.2	49.2	8.6	44.6	344.0	25.5	16.7	48.1
Minimum elevation (m)	266	229	216	225	195	140	61	320	33
Maximum elevation (m)	860	886	1114	796	1117	1245	351	981	354
Elevation range (m)	594	657	898	571	923	1105	290	661	321
Stream order	4th	4th	4th	3rd	4th	5th	3rd	3rd	5th
Drainage density (km/km^2)	1.95	1.77	2.16	2.45	2.38	0.97	1.81	1.57	2.28
% Forested land	76.2	77.6	89.2	66.7	92.2	85.5	39.3	95.1	40.5
% Developed land	8.3	12.7	1.5	16.6	1.0	4.7	26.5	0.8	7.9
% Agricultural land	14.6	8.8	7.0	15.5	5.6	8.0	28.6	0.6	44.8
% Other land	1.7	0.7	0.8	2.1	1.1	1.1	5.6	3.5	6.8

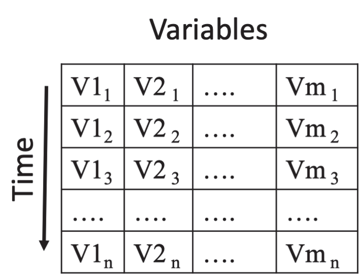


Figure S1: A matrix representation of multivariate time series (m variables, n time steps); a column for each variable and a row for variable value at each time step.

Algorithm K-medoids

Input: storm events (i.e., their multivariate time series representations); number k of clusters to be generated.

Output: k clusters generated from the events.

Procedure

Randomly select k events as medoids from the input events.

```
while termination criteria are not met do
    // Termination condition can be convergence of medoids or maximum allowed iterations.
    Phase 1: Assign each event to its closest medoid.
    Phase 2: From each cluster consisting of the medoid and events assigned to it, select an event that
        gives the smallest sum of distances to all the other events in the cluster and make the selected
        event a new medoid.
end
Return each cluster, consisting of a medoid and all events assigned to it.
```

Algorithm S1: K-medoids algorithm for hydrological event clustering.

Algorithm DTW

Input: $T1$, $T2$: time series, W: warping window size

Output: distance between $T1$ and $T2$

Procedure

Let a and b be the lengths of $T1$ and $T2$, respectively.

Let m be the number of variables in $T1$ and $T2$, respectively.

Create a distance matrix D of size $a \times b$ and initialize all matrix elements to ∞.

$D[0, 0] := 0$. // Initialize the first entry in D.

$i := 1$. $j = 1$. // Initialize the index of a warping path between $T1$ and $T2$.

while $i \leq a$ and $j \leq b$ **do**

  Calculate the squared Euclidean distance, $\sum_{c=1}^{m}(t1_i^c - t2_j^c)^2$, between the ith item in $T1$ and each of the jth item in $T2$ within the range of $j = [i - W, i + W]$.

  Update $D[i, j]$ to $\sum_{c=1}^{m}(t1_i^c - t2_j^c)^2 + \min\{D[i-1, j], D[i, j-1], D[i-1, j-1]\}$.

  increase i by 1.

end

return $\sqrt{D[a, b]}$.

Algorithm S2: Dynamic time warping algorithm for calculating the distance between two time series.

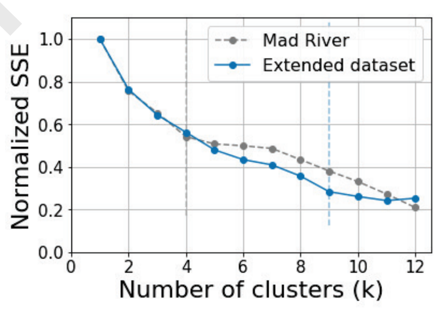


Figure S2: SSE for varying number of clusters for Mad River dataset and Expanded dataset.

Table S2: Default parameter settings for synthetic hydrograph and concentration-graph generator.

Hydrograph					
Type	**Duration-of-peak**	**Time-to-peak**	**Delay**	**Recess**	**Initial Baseflow**
Flashy - early peak return to baseflow	0.4	0.5	0	0.1	0
Flashy - early peak incomplete return to baseflow	0.4	0.5	0	0.4	0
Early peak slow return to baseflow	0.8	0.2	0	0.1	0
Early peak incomplete return to baseflow	0.8	0.2	0	0.4	0
Mid-peak return to baseflow	0.8	0.5	0	0.1	0
Mid-peak incomplete return to baseflow	0.8	0.5	0	0.4	0
Delayed rise to mid-peak return to baseflow	0.8	0.5	0.2	0.1	0.1
Delayed rise to mid-peak incomplete return to baseflow	0.8	0.5	0.2	0.4	0.1
Concentration-graph					
Type	**Duration**	**Time-to-peak**	**Onset**	**Recess**	**Storm-flow**
Early peak	0.5	0.5	0	0	0
Late peak	0.5	0.5	0.5	0	0

Table S3: Result of post-hoc Tukey HSD test for all pairwise comparisons of antecedent conditions metrics. Within each classification scheme if two classes/clusters do not share a letter the mean metric value is significantly different ($alpha = 0.05$).

Antecedent conditions						
Metric	T_{LASTP}	$A3P$	$A14P$	$SM_{SHALLOW}$	SM_{DEEP}	BF_{NORM}
METS clusters						
cluster 1	a	a	a	a	a	a
cluster 2	a	b	b	a	a	b
cluster 3	a	b c	b	a	a	a b
cluster 4	a	c	b	a	a	b
Hysteresis classes						
Class I	a	a b	a	a	a	a
Class II	a	a	a	a	a	a
Class III	a	b	a	a	a	a
Class IV	a	a b	a	a	a	a
Class V	a	a b	a	a	a	a
Complex	a	a b	a	a	a	a

Table S4: Result of post-hoc Tukey HSD test for all pairwise comparisons of rainfall characteristics metrics. Within each classification scheme if two classes/clusters do not share a letter the mean metric value is significantly different ($alpha = 0.05$).

Rainfall characteristics				
Metric	P	P_{MAX}	D_P	T_{PSSC}
METS clusters				
cluster 1	a	a	a	a
cluster 2	b	b	a	b
cluster 3	b	b	a	a
cluster 4	a	c	b	b
Hysteresis classes				
Class I	a b	a	a b	a b
Class II	a	a	a	c
Class III	b	a	a	d
Class IV	a b	a	a b	a
Class V	a b	a	a b	a b
Complex	a b	a	b	b d

Table S5: Result of post-hoc Tukey HSD test for all pairwise comparisons of streamflow and sediment characteristics metrics. Within each classification scheme if two classes/clusters do not share a letter the event metric value is significantly different ($alpha = 0.05$).

Streamflow and sediment characteristics					
Metric	BL	Q_{NORM}	$Log(Q_{NORM})$	SSL_{NORM}	$FLUX_{NORM}$
METS clusters					
cluster 1	a	a	a	a	a
cluster 2	b	b	b	a	a
cluster 3	a c	a b	b	a	a
cluster 4	b c	a b	a	a	a
Hysteresis classes					
Class I	a b	a	a b	a	a
Class II	c	a	a	a	a
Class III	a b	a	b	a	b
Class IV	a c	a	a b	a	a
Class V	a b	a	a b	a	a
Complex	b	a	a b	a	a

Hysteresis class	**Count**
I - Linear (Counter-clockwise)	7
II - Clockwise	0
III - Counter-clockwise	30
IV - Linear then clockwise	0
V - Figure eight	9
Complex (Counter-clockwise)	10
Total	56

Table S6: Distribution of hysteresis loop classes over METS cluster 5 (when $K = 9$) in the expanded dataset ($n = 56$).

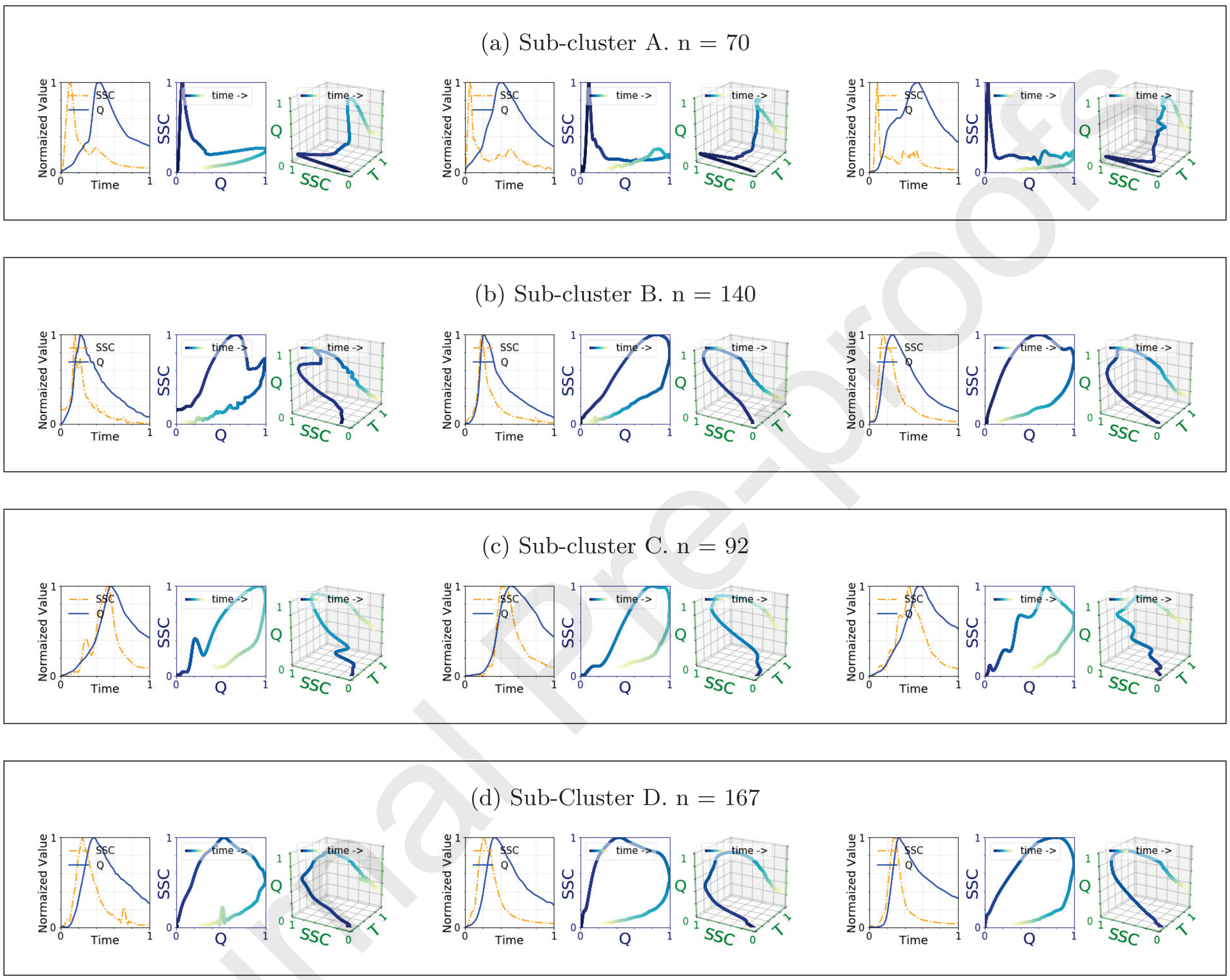


Figure S3: Three storm events closest to the centroid of the four extended dataset tandem clockwise hysteresis sub-clusters ($K = 4$, $N = 496$) — (a) cluster 1 events have sedigraph peaks that occur well before the hydrographs resulting in an “L” shaped loop, (b) cluster 2 have quickly rising hydrographs and sedigraphs, (c) cluster 3 have slow rising hydrographs and sedigraphs, and (d) cluster 4 have sedigraphs that peak before the hydrographs resulting in broad clockwise loops.

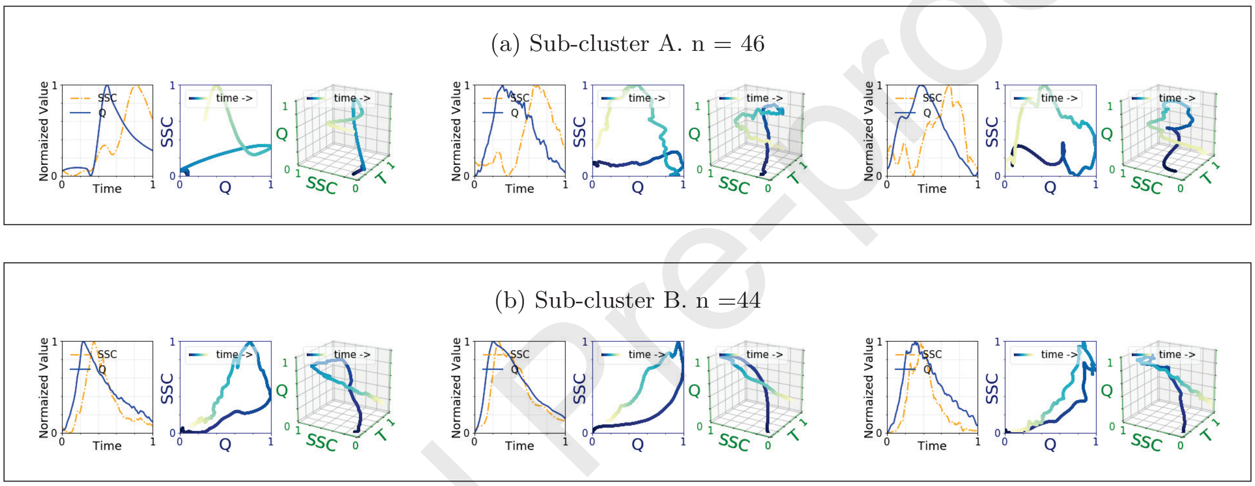


Figure S4: Three storm events closest to the centroid of the four extended dataset tandem counter clockwise hysteresis sub-clusters ($K = 2$, $N = 90$) — (a) cluster 1 events have sedigraph peaks that occur well after the hydrographs resulting in an approximate mirror image of "L" shaped loop and (b) cluster 2 events have sedigraph peaks that occur slightly after the hydrograph peaks.