ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE

10.1029/2021WR030551

Machine-Learning Reveals Equifinality in Drivers of Stream DOC Concentration at Continental Scales

Kristen L. Underwood[1,2], Donna M. Rizzo[1,2], John P. Hanley[3], Gary Sterle[4], Adrian Harpold[4], Thomas Adler[5], Li Li[6], Hang Wen[6], and Julia N. Perdrial[2,5]

[1]Department of Civil and Environmental Engineering, The University of Vermont, Burlington, VT, USA, [2]GUND Institute for Environment, The University of Vermont, Burlington, VT, USA, [3]Department of Microbiology and Molecular Genetics, The University of Vermont, Burlington, VT, USA, [4]Department of Natural Resources and Environmental Science, University of Nevada, Reno, Reno, NV, USA, [5]Department of Geography and Geosciences, The University of Vermont, Burlington, VT, USA, [6]Department of Civil & Environmental Engineering, The Pennsylvania State University, University Park, PA, USA

Key Points:

- Drivers of stream dissolved organic carbon (DOC) concentration exhibit equifinality at the continental scale
- Evolutionary algorithm (EA) identified unique clusters of long-term monitoring sites with elevated stream DOC concentration
- EA identifies specific value ranges of catchment attributes associated with high and low DOC

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

K. L. Underwood,
Kristen.Underwood@uvm.edu

Citation:

Underwood, K. L., Rizzo, D. M., Hanley, J. P., Sterle, G., Harpold, A., Adler, T., et al. (2023). Machine-learning reveals equifinality in drivers of stream DOC concentration at continental scales. *Water Resources Research*, *59*, e2021WR030551. https://doi.org/10.1029/2021WR030551

Received 21 MAR 2022
Accepted 5 MAR 2023

Author Contributions:

Conceptualization: Kristen L. Underwood, Julia N. Perdrial
Data curation: Kristen L. Underwood, Gary Sterle, Adrian Harpold
Formal analysis: Kristen L. Underwood, Donna M. Rizzo, John P. Hanley, Li Li, Julia N. Perdrial
Funding acquisition: Julia N. Perdrial
Methodology: Kristen L. Underwood, Donna M. Rizzo, John P. Hanley, Gary Sterle
Project Administration: Julia N. Perdrial
Resources: Donna M. Rizzo, Adrian Harpold, Julia N. Perdrial

Abstract Research at long-term catchment monitoring sites has generated a great volume, variety, and velocity of data for analysis of stream water chemistry dynamics. To harness the potential of these big data and extract patterns that are indicative of underlying functional relationships, machine learning tools have advantages over traditional statistical methods, and are increasingly being applied for dimension reduction, feature extraction, and trend identification. Still, as examples of complex systems, catchments are characterized by multivariate factor interactions and equifinality that are not easily identified by most machine-learning methods. Using dissolved organic carbon (DOC) dynamics as an illustration, we applied a new evolutionary algorithm (EA) to extract geologic, topographic, meteorologic, hydrologic, and land use attributes that were correlated to mean stream DOC concentration in forested catchments distributed across the continental United States. The EA reduced dimensionality of our attribute dataset to identify the combination of factors, and their specific value ranges, that interacted to drive membership in High or Low mean DOC clusters. High mean DOC concentrations were associated with two distinct geographic locations of variable climatic and vegetative conditions, indicating equifinality. Our findings underscore the importance of critical zone structure in mediating hydrological and biogeochemical processes to govern DOC dynamics at the catchment scale. This multi-scale, pattern-to-process approach is being applied to refine hypotheses for process-based modeling of DOC dynamics in forested headwater streams at catchment to site scales.

1. Introduction

Collaborative research and long-term monitoring of environmental processes in catchments are generating a high volume and variety of data of increasing velocity and veracity (Demchenko et al., 2013; Reichstein et al., 2019), holding the potential to characterize complexity across scales and refine hypotheses on catchment-scale ecosystem response to environmental change. Especially with the advent of new technology such as sensors, lidar, and satellite imagery, we are amassing large amounts of data that capture spatially and temporally variable properties, patterns and processes characterizing the Earth system. Continental-scale data sets of catchment attributes have been compiled to facilitate investigations, such as the Catchment Attributes and MEteorology for Large-sample Studies (CAMELS) data set (Addor et al., 2017; Newman et al., 2015). Similarly, continental scale water quality databases are becoming available such as a data set for German catchments (Musolff, 2020) and a newly assembled relational database for CAMELs catchments (Sterle et al., 2022).

The sheer volume of data, and the considerable variance in data types and temporal and spatial resolution, present a challenge for data analysis using traditional statistical methods (Demchenko et al., 2013). Often parametric and frequentist methods are not robust to conditions that are common in large data sets, including occurrences of missing data, censored data, mixed data types (continuous, nominal, ordinal) or distribution types (e.g., Gaussian, log-Normal, gamma, Poisson). Null-hypothesis significance testing applied to increasing numbers of observations is also challenged by decreasing *p* values (Lin et al., 2013). Moreover, in complex systems, multivariate complexities arise that are not well captured by frequentist statistical methods. Such multivariate complexity may include at least two phenomena: factor interactions and heterogeneity (Hanley, Rizzo, Buzas, & Eppstein, 2020; Urbanowicz et al., 2013). Here we use the term "factor interactions" to refer to a condition when a combination of variables may interact, synergistically or adversely, to result in an effect that neither variable would cause

Software: John P. Hanley
Supervision: Julia N. Perdrial
Validation: Kristen L. Underwood
Visualization: Kristen L. Underwood, Donna M. Rizzo, John P. Hanley, Li Li, Julia N. Perdrial
Writing – original draft: Kristen L. Underwood, Donna M. Rizzo, John P. Hanley, Gary Sterle, Adrian Harpold, Li Li, Julia N. Perdrial
Writing – review & editing: Kristen L. Underwood, Donna M. Rizzo, John P. Hanley, Adrian Harpold, Thomas Adler, Li Li, Hang Wen, Julia N. Perdrial

independently to the same degree. An example of factor interactions from the field of medicine would be when a cancer outcome does not manifest unless there occurs a combination of risk factors (e.g., a genetic predisposition and environmental stressor). Similarly, in catchment science, we are seeking to identify combinations of factors (e.g., catchment attributes) associated with a particular outcome (e.g., constituent response in streams), but which individually may have little or no influence. Thus, we are searching for combinations of factors joined by logical AND operators. "Heterogeneity" is a related property of complex systems, whereby independent sets of variables (some possibly with factor interactions) result in a similar outcome (Hanley, Rizzo, Stevens, et al., 2020; Urbanowicz et al., 2013). In practice, heterogeneity is defined by factors and/or conjunctive clauses joined by logical OR operators, and thus we may require multiple models to comprehensively predict a given outcome (Hanley, Rizzo, Stevens, et al., 2020). In the water resources and ecological fields, heterogeneity is more commonly labeled as "equifinality," to describe a given outcome or response variable (e.g., water quality condition) that may result independently from different combinations of stochastically generated explanatory variables (Beven, 1993). Therefore, we will use the term, equifinality, for the remainder of this paper.

Identifying these factor interactions in catchment science is challenging because many parametric statistical methods are designed to identify only a single *best* model. In addition, despite painstaking efforts to reduce the number of explanatory variables, the search space that results when all factors are combined with their possible ranges of values quickly becomes larger than is possible to exhaustively search with most statistical techniques and today's computational power. As a result, the aim is to find the most parsimonious model (i.e., model that uses the fewest factors to explain an outcome). Another computational challenge encountered when attempting to increase the statistical power of traditional multivariate models is that ordinal and continuous-valued factors often need to be reduced (e.g., binned a priori by domain experts). This data reduction often adds bias and obfuscates important relationships between factors and the outcome class (Bustamante et al., 2014; Murray et al., 2012).

Machine-learning tools hold promise for the simultaneous detection (and characterization) of factor interactions and equifinality that has posed a challenge for more traditional statistical methods. Machine-learning tools are increasingly used for dimension reduction, trend identification, and feature extraction in Critical Zone science (Reichstein et al., 2019; Shen et al., 2018), yet factor interactions and equifinality have been less studied. Because equifinality is a characteristic of many complex systems, not just catchments or ecosystems, it can be valuable to look to other disciplines for their applications of new machine-learning approaches. Some studies from the medical fields have employed Random Forest (RF) approaches to examine data sets for equifinality (e.g.,Goldstein & Rigdon, 2019). Yet, other studies have identified that RFs struggle to identify equifinality (Hanley, Rizzo, Buzas, & Eppstein, 2020). The authors have developed an age-layered, tandem evolutionary algorithm (EA) that has successfully detected factor interactions and equifinality when applied to large data sets for original applications in epidemiology (Hanley, Rizzo, Buzas, & Eppstein, 2020). This machine-learning method is well suited to very large data sets where logistic regression methods can fail to exhaustively search all factor combinations and identify factor interactions (Anderson et al., 2020). Following validation on synthetic benchmark data sets that incorporated equifinality (Hanley, Rizzo, Buzas, & Eppstein, 2020), the EA was applied to mine data from a large socioeconomic survey aimed at identifying the drivers of household infestation with an insect that transmits Chagas disease, a mortal condition if left untreated (Hanley, Rizzo, Stevens, et al., 2020). The EA successfully parsed factor interactions and equifinality in variables that led to an increased risk of Chagas disease transmission (Hanley, Rizzo, Stevens, et al., 2020). Motivated by this early success, the algorithm has since been applied in the engineering fields to define equifinality in catchment-scale and reach-scale attributes associated with bridge damage outcomes in an extreme Vermont flood event (Anderson et al., 2020). Our intention in this current work is to introduce the Tandem EA in the water resources and ecology domains and apply it to a lesser studied aspect of large-scale patterns of the continental US (CONUS), that is, to extract factors and examine factor interactions and possible equifinality in the linkage between catchment-scale attributes and water chemistry outcomes, facilitated by the availability of the new CAMELS-Chem data set (Sterle et al., 2022).

As a test case, we focus on the connection between catchment attributes and mean dissolved organic carbon (DOC) in streams. DOC transported by rivers constitutes an important carbon (C) flux in the global C cycle (Aufdenkampe et al., 2011; Perdrial et al., 2014; Schlesinger & Melack, 1981) where increases in stream DOC potentially contribute to rising atmospheric CO_2 levels and threaten water quality (Butman & Raymond, 2011; Doctor et al., 2008; Öquist et al., 2014, 2009; Raymond et al., 2013). Thus, DOC originating from forested headwater catchments is monitored across the globe (MacDonald & Coe, 2007). For example, investigations of DOC have connected temporal dynamics in forested streams across the northern hemisphere (Monteith et al., 2007;

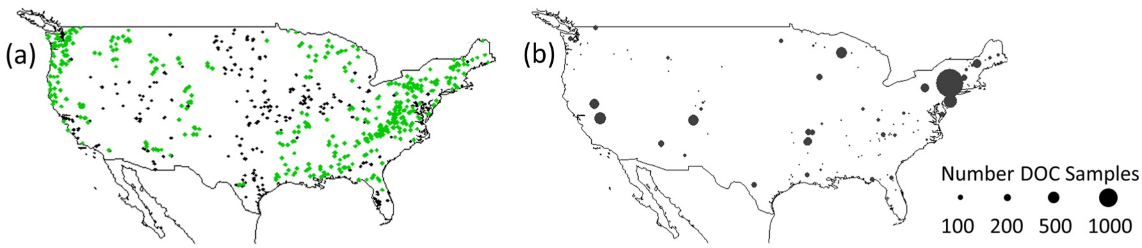

Figure 1. Original Catchment Attributes and MEteorology for Large-sample Studies (CAMELS) sites (n = 671) reduced to (a) n = 449 catchments with forest cover greater than 50% (noted in green); (b) number of available dissolved organic carbon (DOC) observations at 134 of these stations is denoted by a symbol size equal to the square root (x)/10. See also Figure S1 in Supporting Information S1.

Porcal et al., 2009) to a number of drivers, including changing temperature or hydroclimatic conditions (Eimers et al., 2008; Freeman et al., 2001; Lepistö et al., 2008; Worrall & Burt, 2007), changes in deposition (De Wit et al., 2007; Evans & Monteith, 2001; Findlay, 2005; Hruška et al., 2009; Monteith et al., 2007), changes in the ionic strength or redox conditions in soils (Knorr, 2013; Steele & Aitkenhead-Peterson, 2012), and changes in vegetation cover (Finstad et al., 2016) or land management practices (Yallop & Clutterbuck, 2009). One study began to connect catchment attributes using CAMELS-Chem to the directionality of DOC trends in streams in the NE US (Adler et al., 2021), and while variability was high, catchment attributes seemed to partially impact directionality of DOC trends. However, before exploring more temporal patterns, we believe establishing the connection of catchment attributes to mean DOC concentration in streams is helpful as a baseline against which further studies can investigate for potential shifts in these dynamics. We therefore apply machine learning to investigate linkages between catchment attributes and mean DOC concentrations. Specifically, we ask: Can our novel EA identify combinations of catchment attributes that interact to result in spatial patterns of DOC response? Is a similar stream DOC response linked to different combinations of catchment attributes at the continental scale, indicating equifinality?

2. Methods

2.1. Study Area

Our study area comprises the contiguous United States (CONUS) and relies upon the CAMELS data set (Addor et al., 2017; Newman et al., 2015) compiled for 671 catchments of the US Geological Survey (USGS) National Water Information System (NWIS) (Figure 1a). These are minimally disturbed catchments filtered from the Hydro-Climatic Data Network (Lins, 2012) and represent a full range of ecoregions (Omernik, 1987). From the CAMELS catchment attributes that span categories of topography, climatic indices, hydrological signatures, land cover characteristics, soil characteristics, and geological characteristics, we selected 54 catchment attributes with a possible relationship to riverine DOC export (Table S1 in Supporting Information S1). For clarity, we have retained the same variable abbreviations used in Addor et al. (2017).

2.2. Site Selection and Data Harvesting

We further controlled for land cover by sub-setting these sites to 449 catchments with forest cover greater than 50% of the catchment area (Figure 1a). To complement the CAMELS data set, we harvested available water quality data from the USGS NWIS (n = 593) (Sterle et al., 2022). The sampling period ranged from water years 1898 through 2018 comprising 2,695,401 unique records. We summarized mean concentration data and number of observations for pH, temperature, and DOC (Figure 1b).

2.3. Site Clustering

To generate categorical outcome classes required by the Tandem EA, we clustered our CAMELS sites using two approaches and tested for factor interactions and equifinality in catchment attributes linked to these catchment subgroupings. We first defined subgroupings of catchments based on the 54 catchment attributes (Table S1 in Supporting Information S1) relying on hierarchical agglomerative clustering. We clustered all 449 forested catchment sites into groups with similar hydrologic, geological, topographic and vegetative conditions that have a potential relationship to stream DOC efflux. Clustering was performed using the nonparametric hierarchical agglomerative clustering method with Ward's linkage criterion (Ward, 1963). Values were imputed for a very few observations with missing data (i.e., *root_depth*-50 (4.7%), *root_depth*_99 (4.7%), *geol_porosity* (0.2%)). Catchment-attribute values were standardized prior to analysis by subtracting the mean and dividing by the standard deviation. The number of clusters (k) was chosen based on the break-point in the plot of cluster-separation distance and cluster number (i.e., as distance between clusters became marginal with an increase in cluster number). Hierarchical clustering was performed using JMP software (JMP, 2019). In a second approach, we defined catchment subgroupings based on mean DOC concentration. For the subset of CAMELS sites where ≥3 observations of DOC were available ($n = 91$; Figure S1 in Supporting Information S1), each site was identified as having high or low mean DOC concentration using a Jenks natural breaks classification (Jenks, 1967) applied to log-transformed data. Analysis and plotting were performed using the R programming language (R Core Team, 2019) and the "BAMMtools" package (Rabosky et al., 2014). For each approach, we then used the labeled clusters as outcome classes in our factor-selection algorithm to search the multidimensional data space and select catchment attributes that were most strongly correlated with the outcome classes.

2.4. Factor Selection

To identify factors (i.e., most important catchment attributes) associated with a target outcome (i.e., cluster identity or high/low DOC concentration), we used a new, tandem EA with particular advantages for factor selection in large observational data sets of complex systems (Hanley, Rizzo, Buzas, & Eppstein, 2020). Details of the algorithm and its validation on benchmark data sets are provided in Hanley, Rizzo, Buzas, and Eppstein (2020). The motivation for using the tandem EA might be similar to a logistic regression application, in that we can model the probability that catchments with certain attributes belong to a given outcome class. Both methods can model attribute linkages to binary outcome classes (e.g., high or low mean DOC concentration) or to multiple outcome classes as long as they are categorical (e.g., cluster assignment); this latter application would be analogous to the multinomial generalization of logistic regression. However, the EA has several advantages over logistic regression that we exploit in applying factor selection to our CAMELs-Chem data set. First, the EA can be applied to nonparametric data and is robust to varying data types (nominal, ordinal, continuous), skewed distributions, bounded data (e.g., proportions that are bounded between values of 0 and 1 such as the percent organic fraction variable in Table S1 in Supporting Information S1), censored data (e.g., water quality data that have a minimum or maximum reporting limit), and missing values (Anderson et al., 2020; Hanley, Rizzo, Stevens, et al., 2020). The EA is also suitable for exploring data where the number of observations in each target class (or cluster) is unequal (Hanley, Rizzo, Buzas, & Eppstein, 2020).

A second advantage of the tandem EA is that it finds interactions between multiple variables (e.g., catchment attributes) that may result from either additive processes or factor interactions. In two stages, the algorithm identifies and archives two types of clauses that are below a given threshold. First, a conjunctive clause (CC) is a combination of variables that may or may not be correlated and somehow interact to produce an outcome; this sequence of variables is joined by the word "AND." For example, occurrence of a high-magnitude flood (i.e., the outcome) may be associated with catchment-scale attributes such as steep slopes AND shallow soils AND intense rainfall. A CC is often multi-order (i.e., comprises two or more variables) but can also be first-order (i.e., univariate). The previous example would be considered a third-order CC. Another type of clause identified in the second stage of the tandem EA consists of a sequence of CCs that are linked with a logical "OR" statement. This construct is formally known as a disjunctive normal form clause; for simplicity, we will refer to this as a disjunctive clause (DC). DCs are multi-order, while the CCs comprising a DC can themselves range from first-to multi-order. DCs are a particular strength of this novel algorithm, as their identification suggests equifinality in data sets, where the same outcome class may result from different combinations of variables. For example, a high-magnitude flood may result from: (high antecedent soil moisture + rainfall) OR (steep slopes + shallow soils + intense rainfall) OR (thick snow pack + high temperatures). Thus, the EA identifies sets of multivariate interactions (CCs) with a high probability of being associated with an outcome class. In this respect, there are multiple solutions with varying degrees of precision, sensitivity and specificity, making the tandem EA a very appropriate tool for complex systems to identify suites of independent predictor variables (e.g., factors) for each class outcome, especially where equifinality may exist (Hanley, Rizzo, Stevens, et al., 2020).

A third advantage of the tandem EA, is that it not only extracts factors significantly associated with given outcome classes, but also identifies the specific value ranges associated with those factors (Anderson et al., 2020; Hanley,

10.1029/2021WR030551

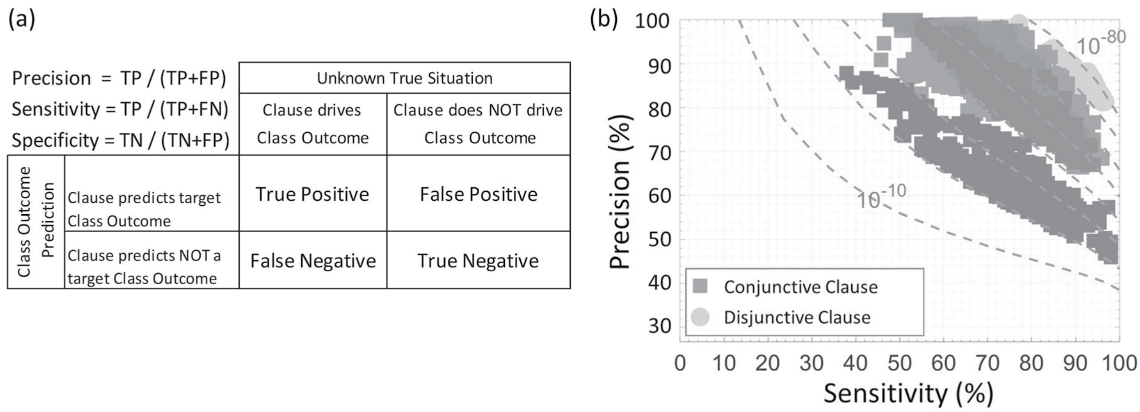

Figure 2. (a) Confusion matrix of binary predictions for each target class. The goal is to maximize the number of true positives (TPs) and true negatives (TNs), while simultaneously minimizing the number of false negatives (FNs) and false positives (FPs). (b) Hypothetical plot of evolved clauses and their precision versus sensitivity.

Rizzo, Buzas, & Eppstein, 2020). In the above hypothetical example of a CC linked to high-magnitude floods, the EA will evolve a specific value range based on the input observations: (catchment slope [2.3, 11%] + soil depth [0.1, 1.3 m] + rainfall intensity [7.6, 43 mm/hr]). In contrast, a low outcome class for flood magnitude could be associated with different variable ranges: (catchment slope [0.01, 2.2%] + soil depth [1.4, 4.8 m] + rainfall intensity [1, 7.6 mm/hr]).

2.4.1. Evolutionary Algorithm

The inputs to the EA consist of an $n \times p$ array where n represents the number of observations and p is the number of variables. The associated outcome class for these observation data must be categorical and the number of target outcome classes, k, must be provided to the algorithm. We can use continuous data as an outcome provided that we first bin the data into meaningful ranges or categories. The EA operates as a sort of binary classifier for each of k user-defined outcome classes, and the algorithm-assigned class outcomes are then compared to the expert-assigned labels to examine the rate of true positives. A binary classifier generates four possible outcomes (Figure 2a). A *true positive* occurs when the algorithm predicts a positive outcome class and the actual value is also positive. However, when the actual value is negative, a positive prediction would be labeled *false positive*. When the actual value is negative, and the prediction is negative, it is termed a *true negative*; a negative prediction when the actual value is positive results in a *false negative*.

For each outcome class, a graphical plot is generated of evolved clauses based on their precision versus sensitivity. Precision (also called positive predictive value) is a ratio of true positives to the sum of true and false positives, expressed as a percentage. For example, in our case of mean DOC outcome classes, a positive outcome would be defined when a given catchment actually exhibits a high mean DOC concentration and the EA correctly predicts high DOC based on a single factor or combination of factors. Sensitivity (also known as true positive rate) is a ratio of true positives to the sum of true positives and *false negatives*. An EA outcome that correctly identifies a high proportion of actual positives will have a high true positive rate (or high "recall," to use a term from the machine-learning disciplines). We will use the term "sensitivity," consistent with language most often used in the medical disciplines. The diagnostic plot (Figure 2b) is a graphical representation of the trade-offs between precision and sensitivity of the EA to predict a given clause's association with a given outcome class. Many solutions are possible, and the best solutions will occupy the upper-right-hand corner of this plot, where both precision and sensitivity are maximized. We avoid those solutions that are positioned mid-way along the top or right axes of the plot. An outcome at 100% precision would mean that we have successfully minimized Type I errors (i.e., a low occurrence of false positives), but this outcome in the absence of high sensitivity could also mean that we have overfit our data. An outcome at 100% sensitivity would mean that we have successfully minimized Type II errors (i.e., a low occurrence of false negatives), but we would sacrifice precision of the test, since we would fail to identify some of the true positives. Instead, we focus on unique CC and DC solutions that occupy the "knee" point along the nondominated pareto front (Das, 1999). In our application, this knee represents the best compromise solution between precision and sensitivity and occupies the region closest to the upper right corner of Figure 2b. By reviewing the individual confusion matrices for those CCs and DCs, we then select clauses that maximize the true positive rate, while simultaneously minimizing the occurrence of false positives and false negatives. Effectively, this process selects for a clause with maximum specificity—that is, correctly identifying as negative a large proportion of those actual negatives. A rule or rule set (CC or DC) with both high sensitivity and high specificity can be described as highly credible.

As the EA evolves, it only retains (i.e., archives) a subset of plausible combinations that exceed a user-specified fitness threshold, which represents a balance between the acceptance rates of false positives and false negatives. This threshold is somewhat analogous to a level of significance in frequentist statistics, except that the frequentist *p*-value is derived from a cumulative distribution function, while the tandem EA uses a hypergeometric probability mass function (PMF) to calculate fitness of a given CC or DC. We view this fitness score as a continuous form of *p*-value that moves beyond the dichotomous *p*-value of frequentist approaches (Wasserstein et al., 2019). The hypergeometric PMF is used as an objective (or fitness) function to quantify the likelihood that the observed association between the CC or DC (unique combination of variables) and the target outcome (e.g., High or Low DOC) is due to random chance. Fitness of a given clause is calculated using the hypergeometric PMF, as:

$$\text{Fitness of clause} = \frac{\begin{bmatrix} X_\mathrm{T} \\ x_\mathrm{M} \end{bmatrix}\begin{bmatrix} N_\mathrm{T} - X_\mathrm{T} \\ n_\mathrm{M} - x_\mathrm{M} \end{bmatrix}}{\begin{bmatrix} N_\mathrm{T} \\ n_\mathrm{M} \end{bmatrix}}$$

where

- N_T is the total number of observations in the data set;
- X_T is the total number of those observations in the target outcome class, k;
- n_M is the total number of observations (i.e., variables) whose features match a given clause; and
- x_M is the number of observations that match the clause and are in target outcome class, k.

Thus, a very low fitness score indicates a very low probability that the clause has been identified purely by chance—in other words, there is a high probability that the clause is associated with the target outcome class.

2.4.2. Tandem EA Application

Practically speaking, the EA is implemented in two runs: a first run evolves and archives the CCs, while the second run evolves and archives DCs. Thus, we refer to the application as a tandem EA, and results for these two stages of the EA will be abbreviated as CCEA and DCEA, respectively. In addition to the fitness threshold, the CCEA employs a dynamically adjusted, order-specific feature sensitivity threshold test to prevent archiving of clauses with unwarranted complexity and thus helps to prevent overfitting of the data (Hanley, Rizzo, Buzas, & Eppstein, 2020).

Our research involved two separate applications of the tandem EA for factor selection. First, we applied the tandem EA to 54 catchment attributes inferred to have importance to DOC dynamics. Combinations of these catchment attributes were identified in CCs and DCs with high probability to be linked to one of several cluster assignments obtained when using hierarchical clustering of these same attributes for the 449 forested catchments. In our second application of the tandem EA, these same 54 catchment attributes were examined instead for their possible linkage to an outcome class of High or Low mean DOC concentration using Jenks natural breaks for 91 catchments with sufficient (≥3) observations of DOC in stream water to calculate a mean value. Thus, we have applied the EA to cases of multiple categorical outcomes, analogous to multinomial logistic regression. Computation of the EA was performed in the Matlab programming language (MATLAB, 2018).

3. Results

3.1. Geographical Clustering of Catchments

Based on inputs of 54 catchment attributes (excluding DOC concentration data), the hierarchical agglomerative clustering algorithm identified four separate clusters of CAMELS sites. The clusters show a notable geographic distribution across the CONUS (Figure 3a) and are characterized by unique ranges of catchment attribute

10.1029/2021WR030551

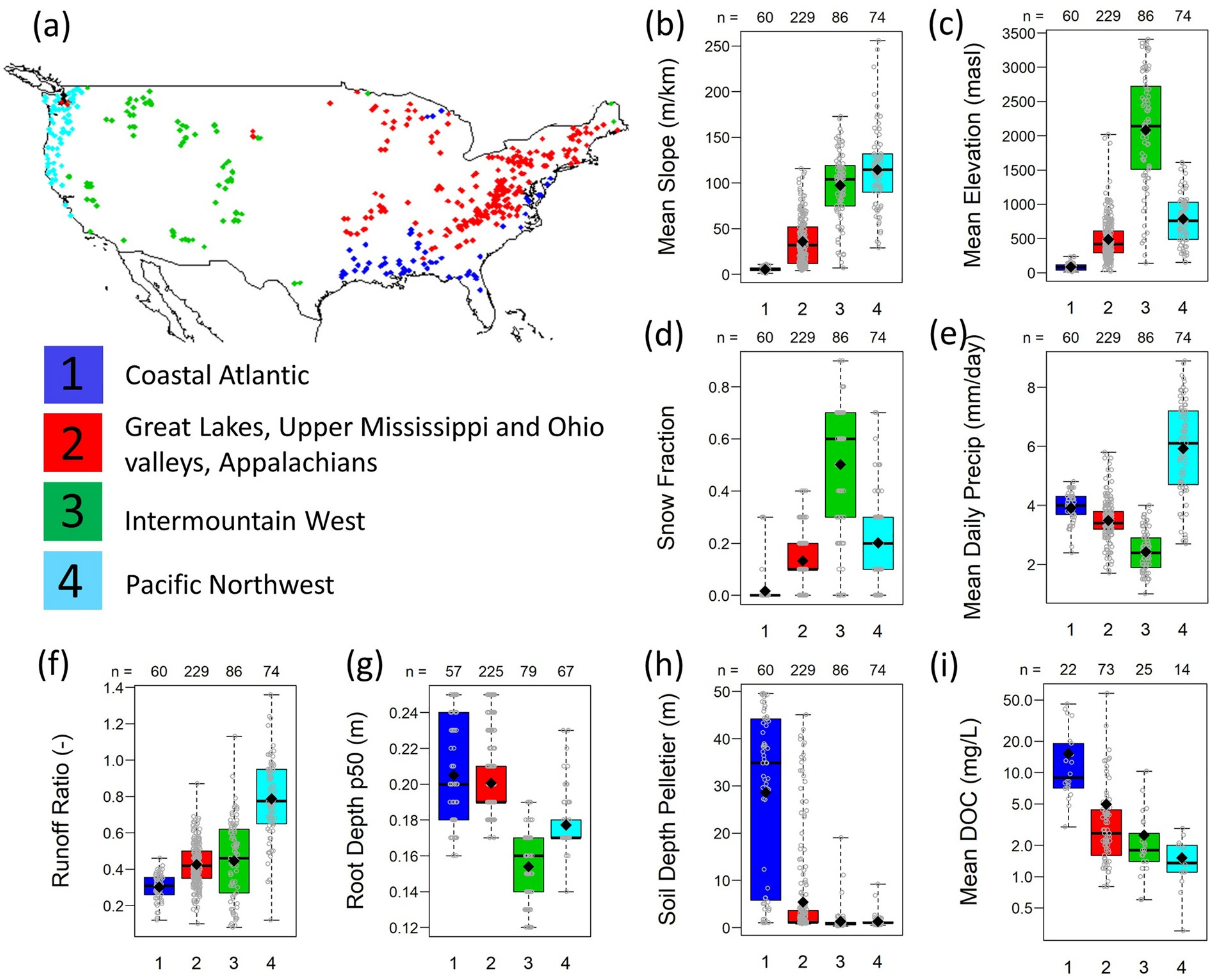

Figure 3. Among the 449 study sites, (a) four geographically dispersed clusters were identified by Hierarchical Agglomerative Clustering on 54 catchment characteristics listed in Table S1 in Supporting Information S1, including factors illustrated in panels (b)–(h). Mean dissolved organic carbon (DOC) concentration in stream efflux (i) was not used as an input to clustering but demonstrates a consistent range across clusters.

variables (Figures 3b–3d) that can be described broadly as topographic, climatic, hydrologic, geologic and land cover factors (Table S1 and Figure S2 in Supporting Information S1).

The tandem EA results provided further insights into specific univariate and multivariate combinations of factors (e.g., specific topographic, climatic, and geologic variables) with highest significance in driving membership in a given cluster, particularly those (Clusters 1 and 2) that span multiple ecoregions. Table 1 provides a summary of the selected (best-fitness) clauses of all those archived by the CCEA and DCEA for each outcome class. For Clusters 1 and 2, respectively, only 106 and 9 DCs passed the fitness threshold (see Section 2.4.2) and were archived by the DCEA. However, for Clusters 3 and 4, no DCs passed the fitness threshold, and therefore were no more informative than the CCs identified during stage 1 of the EA. Thus, catchment groupings in all four clusters were dominated by multivariate factor interactions.

3.1.1. Cluster 1 Outcome Class

Cluster 1 sites comprise coastal locations along the Atlantic Ocean and Gulf of Mexico and are characterized by specific topographic indices (very low elevation and mean slope) and climatic indices (near-zero fraction of precipitation falling as snow) (Figures 3b–3d). Tandem EA results served to reduce dimensionality of the multiple catchment attributes and select factors (catchment variables and their specific value ranges) that were most important in driving Cluster 1 membership. The CC with the best fitness score (Table 1, Figure 4a, Figure S2a in Supporting Information S1) was a fourth-order rule that identified catchments receiving a near-zero fraction of precipitation as snow [*frac_snow*: 0, 0.1], exhibiting thick soil development [*soil_depth_statsgo*: 1.4, 1.5 m]

Table 1
Results of Tandem Evolutionary Algorithm (EA) for Various Outcome Classes

Outcome class	No. archived clauses	Selected clause and order	Fitness	Precision	Sensitivity	Specificity
Geographic clusters						
Cluster 1	1647 CCs	CC 784 (4)	−63.84	0.95	0.95	0.99
		CC 1580 (2)	−61.57	0.92	0.95	0.99
	106 DCs	DC 20 (2)	−52.76	0.78	0.95	0.96
Cluster 2	3305 CCs	CC 783 (3)	−111.9	0.99	0.96	0.99
	9 DCs	DC 1 (2)	−86.7	0.93	0.93	0.93
Cluster 3	2553 CCs	CC 1483 (4)	−66.29	0.79	0.92	0.94
Cluster 4	2627 CCs	CC 2156 (2)	−62.09	0.96	0.86	0.99
Mean DOC concentration in stream water						
High DOC	544 DCs	DC 79 (4)	−19.3	0.97	0.93	0.98
	449 CCs	CC 349 (1)	−15.0	0.87	0.90	0.93
Low DOC	746 DCs	DC 48 (3)	−19.4	0.95	1.0	0.90
	695 CCs	CC 299 (1)	−15.0	0.95	0.93	0.90

Note. Greater negative value for fitness score indicates a better fitness.

in overburden sediments with moderate to high sand fraction [*sand_frac*: 26, 86%], and higher than average mean porosity [*geol_porosity*: 0.16, 0.28]. This CC has a very good fitness score (i.e., large-magnitude negative number), meaning that the probability of this rule being identified merely by chance is very low.

DCs in general for this cluster had a lower precision and specificity than the CCs; note the position of the best-fitness DC (star symbol) in Figure 4a. Thus, many CCs would have greater power to predict Cluster 1 membership than DCs, signifying that Cluster 1 membership is dominated by univariate to multivariate effects. The DC with best fitness and maximum specificity, was a second-order clause comprising a third-order CC and a second-order CC. The first rule identified catchments receiving a near-zero fraction of precipitation as snow [0 to 0.1] falling on soils with moderate to high maximum water content [0.52, 1.05 m], and higher than average mean subsurface porosity [0.16, 0.28]. Alternatively, the second rule of this DC identified catchments with precipitation that is evenly distributed throughout the year [*p_seasonality*: −0.2, 0.2] and low values for the differential between maximum and minimum monthly green vegetative fraction [*gvf_diff*: 0.1, 0.32]. Thus, the DC evolved factors and value ranges that were consistent with (and sometimes the same as) the above CC rule. However, in this case, the CC is a more parsimonious rule, and results in fewer false positives and false negatives than the best-fit DC.

In both the CCEA and DCEA, three outlier stations located along the Great Lakes in Michigan (USGS Stations #4045500, 4056500, and 4057510) were identified as false negatives. These three observations are less similar to other catchments in their cluster given that their dominant geologic class is carbonate sediment rocks, and they have a measurable (although low) fraction of precipitation falling as snow (0.3).

3.1.2. Cluster 2 Outcome Class

Cluster 2 sites include catchments located along the Appalachian Mountain range and upper mid-west, characterized by low-to-moderate mean slope and elevation (Figure 3). Compared to Cluster 1, these catchments receive a somewhat greater percentage of their annual precipitation as snow, given their generally higher elevation and latitude. The CC with best fitness and highest specificity for Cluster 2 (Table 1; Figure 4b, Figure S3c in Supporting Information S1) was a third-order rule that identified catchments with precipitation ranging from evenly distributed through the year to summer-dominated [*p_seasonality*: −0.2, 0.7] and a relatively wide range for mean high-flow duration [*high_q_dur*: 1, 5 days]. Dominantly, these catchments are underlain by carbonate and mixed sedimentary rocks or unconsolidated sediments. The DC for Cluster 2 that exhibited best fitness, while also minimizing the number of false positives and negatives, was a second-order rule set comprising two univariate CCs (Figure S3d in Supporting Information S1). The first rule identified catchments covered by deciduous broadleaf and mixed forests, with a lesser occurrence of cropland/natural vegetation mosaic. The second rule specified

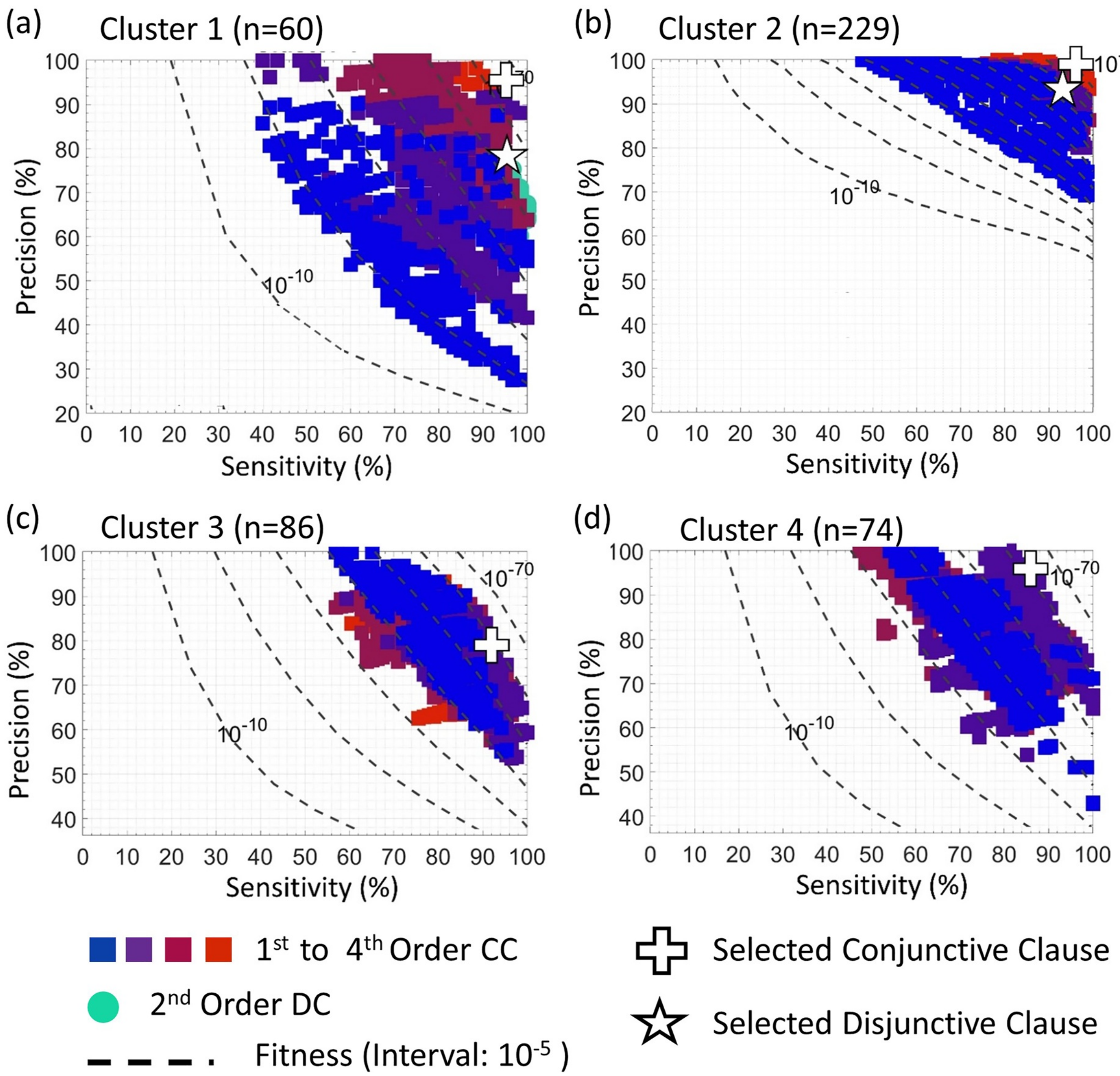

Figure 4. Archived results of the tandem evolutionary algorithm (EA) for each of the four geographical clusters as target outcome classes on a plot of precision versus sensitivity. Each symbol corresponds to an individual outcome of the tandem EA that represents a value range for a combination of catchment attributes (or in the univariate case—a single catchment attribute) with significant probability for driving the class outcome. Conjunctive clauses (CCs) evolved from the CCEA are presented as square symbols, and disjunctive clauses (DCs) evolved from the DCEA are circles, with color hues indicating increasing clause order. The black dashed contour lines represent intervals of equal fitness score with better fitness scores (more negative values) concentrated in the top right corner of the plot (see Section 2.4.1). The tandem EA has reduced dimensionality of the dataset ($p = 54$ attributes) to identify the combination of two to four catchment attributes that interact (synergistically or adversely) and are most significant in driving cluster membership. Geographic cluster membership is dominated by multivariate factor interactions, because many CCs outperform DCs.

a high differential between maximum and minimum average monthly leaf area index [*gvf_diff*: 0.35, 0.55], an indicator of deciduous vegetation.

3.1.3. Cluster 3 Outcome Class

Cluster 3 sites are located predominantly in the intermountain west, and are characterized by very high elevation, steep slopes, comparatively higher fractions of precipitation falling as snow (Figure 3), and less forest cover (Figure S2 in Supporting Information S1). The CC with the best fitness (Table 1; Figure 4c, Figure S3e in Supporting Information S1) was associated with catchments of moderate to high aridity index [0.6, 3], exhibiting relatively shallow overburden depths to bedrock [Pelletier: 0.3, 2.5 m], relatively shallow rooting depths for the upper 50% of the root system [0.12, 0.19 m], and characterized by a variety of dominant land covers, most

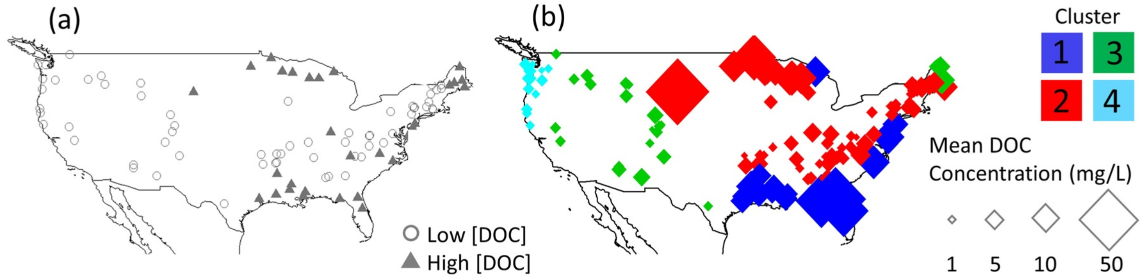

Figure 5. Dissolved organic carbon (DOC) concentration at Catchment Attributes and MEteorology for Large-sample Studies (CAMELs) sites ($n = 91$) with ≥3 observations, classified as Low ($n = 61$) or High ($n = 30$) using Jenks natural breaks classification on log10-transformed mean DOC values (a). For the subset of $n = 133$ sites with available data, mean DOC concentration was internally consistent by assigned cluster (b), where symbol size is graduated by the square root of the DOC concentration. Cluster 1 sites ($n = 22$) had the highest range and mean of DOC concentration.

commonly evergreen needleleaf forest, grasslands, and woody savannas. Using this rule, three distal stations located in Maine and Minnesota (USGS #1013500, 1030500, and 5129115) were identified as false negatives by the CCEA, because they have much deeper overburden depths than the remaining sites (7, 19, 11 m, respectively). Relative to the remaining sites in the cluster, these three outlier stations also have slightly higher values for the differential between maximum and minimum monthly green vegetative fraction (*gvf_diff* and *lai_diff*), a lower snow fraction, lower elevation and mean slope, and lower aridity index.

3.1.4. Cluster 4 Outcome Class

In the Pacific Northwest, Cluster 4 sites are located along the Cascade and northern Coastal Range mountains at low to moderate elevations, with mean slopes in a range similar to catchments of Cluster 3, and snow fractions in a range similar to Cluster 2 (Figure 3). These catchments have comparatively higher mean daily precipitation (occurring dominantly in the winter months), and generally higher runoff ratios than catchments of other clusters (Figure 3). The CC with the best fitness for Cluster 4 (Table 1; Figure 4d, Figure S3f in Supporting Information S1) identified catchments with high mean daily precipitation [3.7, 8.9 mm day^{-1}] and dominant land cover of evergreen needleleaf forests (with minor occurrences of grasslands, mixed forests, and woody savannas).

3.2. Patterns in DOC Concentrations

Mean DOC concentration varied significantly across the CONUS ecoregions and our geographic clusters (Figure 5). The break point between High and Low mean DOC concentration identified by Jenks classification was 4.6 mg/L. Sites with High DOC concentration (>4.6–58 mg/L) were generally located along coastal regions of the eastern CONUS and the Great Lakes, spanning Clusters 1 and 2, while Low DOC concentrations (0.8–4.6 mg/L) were noted for sites along the interior and the west coast, comprising most of Clusters 3 and 4, but also some Cluster 2 stations along the Appalachian Mountains.

Using these two outcome classes (High and Low DOC), the tandem EA identified predictive combinations of catchment-scale physical attributes (e.g., topography, geology, soils, land cover) and hydroclimatic variables with possible importance in driving DOC stream chemistry in our forested catchments. In contrast to the geographic cluster outcomes, the DCEA evolved better-fitness clauses than the CCEA for High and Low DOC outcomes (Table 1, Figure 6), indicating equifinality in the data set.

3.2.1. High Mean DOC Outcome Class

For the High DOC outcome class, two DCs (Figure S4b in Supporting Information S1) were co-located at the knee point (star symbol of Figure 6a), and had equivalent values of precision (96.6%) and sensitivity (93.3%). These two rule sets were composed of the same four univariate to second-order CCs, and they differed only in terms of the component CC for overburden depth having slightly different value ranges: (1) *elev_mean* [40 m, 90 m]; OR (2) *frac_snow* [0.2, 0.3] + *p_mean* [1.9 mm day^{-1}, 3.3 mm day^{-1}]; OR (3) *high_precip_freq* [23 days yr^{-1}]; OR (4) *soil_depth_pelletier* [27.2 m, 44.2–44.6 m]. This *soil_depth_pelletier* catchment attribute expresses the average depth of unconsolidated materials over weathered bedrock to a maximum limit of assessment of 50 m and was developed on a global scale at a 30 arcsec grid (Pelletier et al., 2016) and compiled as the mean of all those ~1 km grid points that fell within each catchment (Addor et al., 2017).

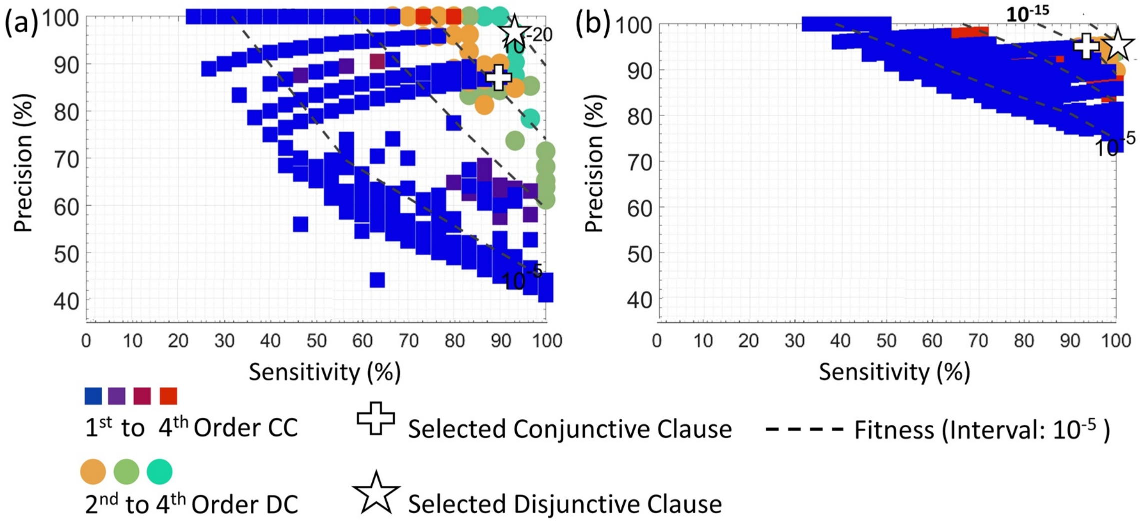

Figure 6. Archived results of the tandem evolutionary algorithm (EA) for each of the target outcome classes: (a) High [dissolved organic carbon (DOC)] ($n = 30$) and (b) Low [DOC] ($n = 61$) on a plot of precision versus sensitivity for ($p = 54$) catchment attributes. Conjunctive clauses (CCs) evolved from the CCEA are presented as square symbols, and the disjunctive clauses (DCs) evolved from the DCEA are plotted as circles, with color indicating clause order. The black dashed contour lines represent intervals of equal fitness score. For each outcome (High DOC, Low DOC), the DCs (stars) outperformed the CCs (plus sign), indicating that the same target outcome results from different predictive combinations of variables—a phenomenon known as equifinality.

If we relied on any one of these component CCs, alone, to predict whether a catchment site had High DOC, they would have very high precision (90%–100%) but quite low sensitivity (27%–50%), because we would fail to accurately classify between 15 and 22 of the 30 catchments in the High DOC class. Instead, taken collectively, these component CCs suggest equifinality as an important quality of the data set (i.e., high mean DOC concentrations can result from different factors or combinations of factors).

Overall, compared to the DCs, the CCs had worse fitness and thus less power to predict High DOC class outcome. The archived CCs were largely univariate in nature—identifying a single catchment attribute instead of a combination of multiple attributes (i.e., factor interactions) as a driver of the High DOC outcome class. Of all the 449 archived CCs, the rule which had near-maximum precision (87.1%) and sensitivity (90%), while simultaneously minimizing false positives (4/91 = 4.4%) and false negatives (3/91 = 3.3%) was overburden thickness [*soil_depth_pelletier*] with a value range from 5 to ≥50 m (Figure S4a in Supporting Information S1).

One catchment in Cluster 2 was identified as a false negative by the best-fitness CC for the High DOC outcome class (Figure S4a in Supporting Information S1; Castle Creek above Deerfield Reservoir near Hill City, SD; USGS Station # 06409000). Three DOC observations recorded at this site (1.2, 2.9, and 170 mg/L) included one possible outlier, which skewed the mean DOC value into the High category (58 mg/L DOC). These results flag this site as one that possibly does not belong to the class it was grouped into, and highlight the relatively low number of observations ($n = 3$) relied upon to classify this site as having High DOC mean concentration.

3.2.2. Low Mean DOC Outcome Class

Compared to the High DOC solutions, the Low DOC solutions had generally higher precision, and again the DCs exhibited better fitness than many of the CCs (Figure 6b). Archived DCEA results for the Low DOC outcome class included several second-to fourth-order DCs, each made up of first to second order CCs. Ten third-order DCs with equivalent fitness scores were co-located at the position indicated by the star symbol on the precision versus sensitivity plot (95% precision; 100% sensitivity; Figure 6b). These third-order rule sets were: (1) *soil_depth_pelletier* [0.6, 1.6 m]; OR (2) *soil_depth_pelletier* [1.1, 4.5 m]; OR (3) *root_depth_50* [0.12, 0.21 m] + *dom_land_cover* [Classes 2, 4, 5, 6, 7, 8]. Among the identified catchment-scale land cover classes, most of the 61 catchments in the Low DOC outcome class were characterized by forest variants, including deciduous broadleaf ($n = 21$), evergreen needleleaf ($n = 16$), and mixed forests ($n = 11$). The value ranges for rooting depth indicate that the top half of the vegetative root system lies on average between 0.12 and 0.21 m, reflecting somewhat more shallow-rooted vegetation in comparison to the full data set (Figure S2 in Supporting Information S1). When two or more variables occur in a rule set, they are correlated to the outcome class but may or

10.1029/2021WR030551

may not be correlated themselves. In this case, *dom_land_cover* and *root_depth_50*, are likely correlated in part because rooting depths were developed using a numerical function tied to land cover class (Addor et al., 2017).

The CC with the best fitness score (Table 1; "plus" sign in Figure 6b) was associated with an overburden thickness [*soil_depth_pelletier*] between 0.4 and 5.4 m, and had a precision of 95.0% and sensitivity of 93.4%. Seventeen of the eighteen lowest-fitness CCs were first-order clauses containing this same variable but slightly different value ranges. This rule had nearly as good a fitness score as the best-fit DC and could be considered a more parsimonious rule for prediction of Low DOC concentrations in stream water with only slightly sacrificed sensitivity (Table 1). The above third-order DC provides more nuance to distinguish between Low DOC catchments in Cluster 2 versus those in Clusters 3 and 4. The same Castle Creek site was identified as a false positive by the best-fitness CC and DC for the Low DOC outcome class (Figures S3c and S3d in Supporting Information S1).

4. Discussion

4.1. Application of Tandem EA for Spatial Pattern Detection in DOC Concentration

We applied a new machine-learning algorithm on an integrative, CONUS-scale "big data" set (CAMELS-Chem) to illustrate data-driven approaches for pattern identification in catchment science. Specifically, we sought to identify factor interactions and equifinality in catchment attributes as drivers of mean DOC response in streams at the continental scale. We first used the tandem EA to reduce dimensionality of our data and identify key catchment attributes at the CONUS scale that are associated with four geographical regions labeled by hierarchical clustering of topographic, geologic, hydrologic climatic and land cover variables (Figures 3 and 4). The four clusters are consistent with ecoregions of Herlihy et al. (2008) that were aggregated from Level III ecoregions (Omernik, 1987), and they mirror the four aggregated geographic regions analyzed for the National Wetland Condition Assessment (USEPA, 2015). In general, geographic clusters were characterized by multivariate factor interactions, but substantial equifinality was not identified.

On the other hand, the tandem EA resolved both factor interactions and equifinality at the CONUS scale in governing variables for streamflow chemistry emanating from these catchments, when DOC concentrations were clustered into high and low mean DOC outcome classes. High DOC mean concentrations (>4.6 m/L) were equal to or higher than the global mean of 5 mg/L (Berner & Berner, 2012), and were associated with two geographic clusters spanning widely variant climatic and vegetative conditions: Cluster 1 catchments located in the Coastal Plains of Southeastern US, and some of the Cluster 2 catchments located along the margins of the Great Lakes in the Upper Midwestern US. These findings confirmed that geography alone does not serve as a skillful predictor for DOC concentrations. Instead, four separate factors, or combinations of factors, made up the best-fit predictive model for High DOC evolved by the tandem EA (right side of Figure 7): (1) low mean elevation (<90 m; predominantly Cluster 1 sites) OR (2) low to moderate values for mean daily precipitation and measurable (although low) annual snow fraction (Cluster 2 and select Cluster 1 and 3 sites) OR (3) a relatively high annual frequency (23 days) of high precipitation days (select Cluster 1 and 2 sites) OR (4) relatively thick overburden materials (>27 m; Cluster 1 and 2 sites). Multiple factors were also associated with the Low DOC outcome class (left side of Figure 7), including (i) very shallow overburden sediments (Cluster 3 and 4 sites) OR (ii) shallow overburden sediments (select Cluster 2 sites) OR (iii) shallow to moderate rooting depths in mixed deciduous broadleaf and evergreen needleleaf forests (Cluster 2, 3, and 4 sites). For complex systems exhibiting factor interactions and equifinality, these multiple models will more comprehensively predict a given outcome than traditional statistical approaches (Hanley, Rizzo, Stevens, et al., 2020). Furthermore, the tandem EA can extract patterns that are meaningful to a given outcome class, even when data on that outcome class are sparse—a condition that exists for many constituents of concern (Sterle et al., 2022). Our selection of DOC was particularly illustrative, since sufficient long-term records were available for only 91 of the 449 catchments included in this analysis. In this sense, sites with limited streamflow chemistry data may benefit from predictive models evolved relying on labeling from other sites that have higher-frequency and longer-term records.

4.2. From Pattern to Process: Integrating Domain Knowledge

Fundamentally, data-driven approaches point us toward important factors and factor interactions, but domain knowledge is required to ascribe possible meaning to them (Goldstein et al., 2018) and refine hypotheses for further testing using process-based methods. Domain knowledge can be particularly helpful to examine and

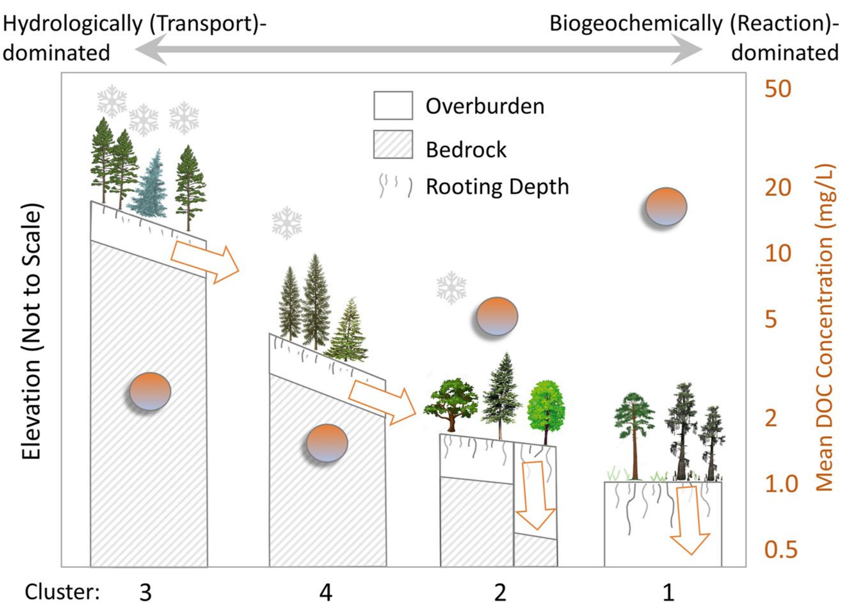

Figure 7. Perceptual diagram of mean dissolved organic carbon (DOC) concentration patterns across the continental US (CONUS). Geographic clusters are arranged on a continuum from high (left) to low (right) elevation, and demonstrate increasing overburden thickness with decreasing elevation. Mean DOC concentration is lesser in high-elevation, high-slope catchments of hydrologically dominated Clusters 3 and 4, and greater in lower-elevation, low-slope catchments of biogeochemically dominated Clusters 2 and 1. The latter groups are also characterized by deeper overburden sediments and rooting depths, supporting more deciduous vegetation.

interpret other lesser-fitness rules or rule sets evolved by the tandem EA, to consider them as a proxy for the best-fit solution, and to interpret possible processes linked to the observed patterns. For example, a nearly equal-probability factor associated with High DOC was the first order CC of deep overburden thickness. This is not a contradiction of the best overall fitness DC but can be considered a consequence of, or covariant with, three of the four clauses comprising the DC. In other words, low elevation regions are predominantly located along coastal margins where sediments have accumulated. A regional climate that produces "wet when warm" conditions leads to high weathering rates and productive vegetation (Perdrial et al., 2015; Rasmussen et al., 2007). Thus, overburden thickness is both an indicator variable of High DOC, but also an integrator variable of these meteorological, hydrologic and soil attributes. Factors correlated with the Low DOC concentrations of intermountain west sites in Cluster three included: shallow overburden thickness (*soil_depth_pelletier*), associated with thin soil development (*soil_depth_statsgo*), high aridity, and shallow rooting depths. To develop overburden thickness estimates on hillslopes, Pelletier et al. (2016) combined erosion rate estimates proportional to topographic curvature and inversely proportional to soil depth and soil production. The success of the Pelletier et al. overburden thickness estimate is perhaps due to its ability to resolve local (i.e., hillslope vs. upland) and regional (i.e., climate, geology, etc.) drivers of critical zone development.

These findings underscore the importance of overburden materials and their structural properties (thickness, porosity, permeability) in mediating hydrological and vegetation processes to govern DOC dynamics at the catchment scale. While the goal of this study was not to provide an in-depth investigation of DOC dynamics across the CONUS, our results are in agreement with current process-based understanding of the balance between biogeochemical and hydrological drivers of DOC dynamics. At the catchment scale, DOC export into streams is typically controlled by the interaction of DOC production, consumption and hydrological transport that is conceptually summarized as ecosystem control points (Bernhardt et al., 2017). High DOC supply is associated with deeper soils on low slopes, and a greater potential for precipitation partitioned to slow and deep groundwater (Cluster 1 and some Cluster 2 catchments, Figure 7), and this expected increased residence time would suggest a more biogeochemically dominated DOC export regime. Low DOC supply, and episodic (flashy) hydrologic connectivity and transport are associated with a greater proportion of streamflow derived from fast-flow pathways and low DOC concentrations overall (Gnann et al., 2019). Similarly, in Cluster 4 (and some Cluster 2) catchments, a transport-dominated condition engendered by moderate to steep slopes but low hillslope curvature, likely are responsible for Low DOC concentrations, *despite* a moderate DOC production capacity—that is, high mean precipitation and mean duration precipitation, and moderate to high soil profile thickness (*soil_depth_statsgo*) in those shallow overburden materials.

While our results largely conform to current understanding of continental-scale DOC patterns, this is actually an outcome that we believe may help to reinforce acceptance of the Tandem EA as a new tool with value for exploratory data analysis applied to less-characterized constituents, or less-studied contexts, such as temporal dynamics. Specifically, these results offer a baseline on the connection of catchment attributes and DOC concentrations in general, against which further studies can test which attributes increase ecosystem resilience/resistance against change.

4.3. Consideration of Uncertainty

To keep the focus on a relatively straightforward application of the novel tandem EA, and to offer a baseline analysis of the connection between DOC and catchment attributes, we have "controlled" for time-transgressive variations in DOC concentration, by aggregating the catchment signal (i.e., relying on an average DOC concentration as the output class), and using annual-scale values for catchment attributes (Addor et al., 2017).

Our work focuses on illustrating a new exploratory data analysis technique capable of detecting equifinality and factor interactions, and does not explicitly address uncertainty of model inputs or outcome classes. Large-sample data sets, often derived from remote sensing sources, can be prone to observation errors (Addor et al., 2020). The CAMELS attributes used as inputs to our Tandem EA represent catchment-averaged values of various hydrologic, geologic, topographic, climatologic and land use variables that do not capture the catchment-specific variability in these properties (Addor et al., 2020, 2017). Uncertainty of these values may be considerable, particularly for subsurface geologic variables derived largely from remote sensing, but estimates of observational uncertainties were not supplied in the original data set. Under-representative sampling may also have biased classification of High and Low mean DOC concentration, used as outcome categories in our algorithm. Of the 449 catchments we clustered into geographic regions, only 91 contained monitoring records with three or more DOC observations; 67% of these 91 catchments had DOC records containing 10 or more observations; and 41% had 30 or more observations (Figure S1 in Supporting Information S1). Yet, notably, the EA flagged a station with sparse outcome data ($N = 3$) as a "false negative," suggesting a degree of method robustness to sources of data uncertainty (Section 3.2.1).

Admittedly, the above-described sources of uncertainty may have led to diminished power to detect factor importance or factor interactions. Yet, issues of observational uncertainty, sparse data, and imbalanced data sets, are common when dealing with large data sets and seeking to extract patterns across broad regions. In light of the limited financial and labor resources available to expand upon these monitoring data sets both in terms of spatial density, monitoring frequency, and record length, we were motivated to develop tools—like the Tandem EA—with greater robustness to these data limitations. A principal value of the Tandem EA is in exploratory data analysis to refine hypotheses for further testing at the regional or catchment scales, where explicit treatment of uncertainty can be incorporated in process-based models and/or data-driven models (e.g., Bayesian hierarchical models).

4.4. Future Work

The tandem EA is transferable to other settings or contexts. Catchments could be categorized in outcome classes for different constituents (e.g., suspended sediment, nitrogen, phosphorus) to explore driving factors for alternate catchment dynamics (weathering, nutrient cycling). The same tool could be applied to data collected at high temporal resolution (e.g., nutrient time series) to suggest possible hourly to seasonal scale drivers of nutrient dynamics, a focus of future research. Given the observed differences in climate stationarity across the CONUS (Hirsch, 2011) and the regional differences in projected trends (Hayhoe et al., 2007; IPCC, 2013), various regions may have been (and may continue to be) exposed to different magnitudes and directionality of hydrologic drivers. From our analysis, we might hypothesize that climatic shifts that continue to promote thick soils may lead to more concentrated stream water DOC in the long term (i.e., conditions that produce thick soils (DOC production) also lead to DOC transport). Future work could apply the same tandem EA approach to examine possible linkages between catchment attributes and temporal trends in DOC concentration. Insights from the current work are being used to inform structure and parameterization of numerical models applied at the site scale at long-term monitoring sites that are representative of these clusters (e.g., CZOs, LTERs). We have identified select catchments representative of these clusters, where greater temporal and spatial resolution of time-series records are available, to carry out process-based modeling of DOC dynamics aided by bench-scale soil experiments (e.g., Adler et al., 2021; Wen et al., 2020). These site-scale models are more computationally and representationally detailed (Larsen et al., 2014), and are being used to elicit processes important in DOC streamflow dynamics from forested headwater streams. Such complementary approaches, combining data-driven models and mechanistic models are better at addressing both "breadth" and "depth" of analysis to advance understanding of biogeophysical processes (Gupta et al., 2014).

5. Conclusions

A novel EA was applied to extract dominant factor interactions and equifinality associated with spatial patterns in stream DOC concentration at predominantly forested, long-term monitoring sites across the continental US. The tandem EA reduced dimensionality of our catchment attribute dataset to identify that the same outcome (High mean DOC concentration, >4.6 mg/L) is associated with different predictive combinations of factors, and their specific value ranges, including (1) low mean elevation, or (2) low to moderate mean daily precipitation with a measurable snow fraction, or (3) relatively high frequency of high-precipitation days, or (4) thick overburden materials. Multiple factors were also associated with the Low DOC outcome class, including (1) very shallow or (2) shallow overburden sediments or (3) shallow to moderate rooting depths in mixed deciduous broadleaf and evergreen needleleaf forests. For complex systems exhibiting such factor interactions and equifinality, these multiple models identified by the EA will more comprehensively predict a given outcome than traditional statistical approaches (Hanley, Rizzo, Stevens, et al., 2020). Because thickness of overburden materials is a factor that may be considered a consequence of, or covariant with, the other identified factors, our findings underscore the importance of critical zone structure in mediating hydrological and biogeochemical processes to govern DOC dynamics at the catchment scale.

Exploratory analysis of big data at broad scales of the CONUS is more inclusive and embraces a fuller diversity of catchment types than can be modeled using deterministic approaches, given the high-resource intensity that such process-based models necessitate. Processes and governing factors important to DOC flux or concentration patterns at this continental scale may (or may not) be significant at a catchment or site scale. Nevertheless, using this data-driven approach, patterns emerge from big data that can be used to shape hypotheses for further study, or refine parameterization using process-based models at the catchment or site scale (Bergen et al., 2019). In turn, findings from a mechanistic approach within a unique catchment can be placed in a broader-scale context by relying on exploratory analysis of big data at the CONUS or global scale. Machine-learning approaches, and specifically the tandem EA, offer important advantages over traditional statistical approaches to identify factor interactions and equifinality that characterize complex systems (Anderson et al., 2020; Hanley, Rizzo, Stevens, et al., 2020). Such complex systems tools can be an important first step in the iterating inductive to deductive steps of a hybrid investigation (Bergen et al., 2019; Larsen et al., 2014; Reichstein et al., 2019).

Acknowledgments

The authors are grateful to the associate editor and two anonymous reviewers for constructive reviews that improved the manuscript. This work was supported by the National Science Foundation under Grant NSF-EAR 1724171 and NSF 2012123, which provided partial support to all authors except JPH who was supported by National Institute of General Medical Sciences of the NIH under Award Number 1P20 GM125498-01. Co-authors DMR and JNP were partially supported by NSF under Vermont EPSCoR Grants EPS-1101317 and NSF OIA 1556770. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, National Institutes of Health, or Vermont EPSCoR.

Data Availability Statement

The data set of catchment attributes was extracted from the CAMELS data set available online (https://doi.org/10.5065/D6G73C3Q), and select water chemistry parameters for forested CAMELS sites are provided in Sterle et al. (2022). Open-source Matlab code for the tandem EA is available at: https://www.mathworks.com/matlabcentral/fileexchange/69950-cceaand-dnfea.

References

Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K., & Mendoza, P. A. (2020). Large-sample hydrology: Recent progress, guidelines for new datasets and grand challenges. *Hydrological Sciences Journal*, *65*(5), 712–725. https://doi.org/10.1080/02626667.2019.1683182

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Adler, T., Underwood, K. L., Rizzo, D., Harpold, A., Sterle, G., Li, L., et al. (2021). Drivers of dissolved organic carbon mobilization from forested headwater catchments: A multi scaled approach. *Frontiers in Water*, *3*, 578608. https://doi.org/10.3389/frwa.2021.578608

Anderson, I., Hanley, J. P., Rizzo, D. M., Huston, D. R., & Dewoolkar, M. M. (2020). Evaluating damage to Vermont bridges by Hurricane Irene with multivariate bridge inspection and stream hydrogeologic data. *Journal of Bridge Engineering*, *25*(10). https://doi.org/10.1061/(ASCE)BE.1943-5592.0001603

Aufdenkampe, A. K., Mayorga, E., Raymond, P. A., Melack, J. M., Doney, S. C., Alin, S. R., et al. (2011). Riverine coupling of biogeochemical cycles between land, oceans, and atmosphere. *Frontiers in Ecology and the Environment*, *9*(1), 53–60. https://doi.org/10.1890/100014

Bergen, K. J., Johnson, P. A., de Hoop, M. V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. *Science*, *363*(6433), eaau0323. https://doi.org/10.1126/science.aau0323

Berner, E. K., & Berner, R. A. (2012). *Global environment: Water, air and geochemical cycles*. Princeton University Press.

Bernhardt, E. S., Blaszczak, J. R., Ficken, C. D., Fork, M. L., Kaiser, K. E., & Seybold, E. C. (2017). Control points in ecosystems: Moving beyond the hot spot hot moment concept. *Ecosystems*, *20*(4), 665–682. https://doi.org/10.1007/s10021-016-0103-y

Beven, K. J. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling. *Advances in Water Resources*, *16*(1), 41–51. https://doi.org/10.1016/0309-1708(93)90028-e

Bustamante, D. M., De Urioste-Stone, S. M., Juárez, J. G., & Pennington, P. M. (2014). Ecological, social and biological risk factors for continued *Trypanosoma cruzi* transmission by *Triatoma dimidiata* in Guatemala. *PLoS One*, *9*(8), e104599. https://doi.org/10.1371/journal.pone.0104599

Butman, D., & Raymond, P. (2011). Significant efflux of carbon dioxide from streams and rivers in the United States. *Nature Geoscience*, *4*(12), 839–842. https://doi.org/10.1038/ngeo1294

Das, I. (1999). On characterizing the 'knee' of the pareto curve based on normal-boundary intersection. *Structural Optimization*, *18*(2–3), 107–115. https://doi.org/10.1007/BF01195985

Demchenko, Y., Grosso, P., de Laat, C., & Membrey, P. (2013). Addressing big data issues in scientific data infrastructure. In *2013 International conference on collaboration technologies and systems (CTS)* (pp. 48–55). https://doi.org/10.1109/CTS.2013.6567203

De Wit, H. A., Mulder, J., Hindar, A., & Hole, L. (2007). Long-term increase in dissolved organic carbon in streamwaters in Norway is response to reduced acid deposition. *Environmental Science & Technology*, *41*(22), 7706–7713. https://doi.org/10.1021/es070557f

Doctor, D. H., Kendall, C., Sebestyen, S. D., Shanley, J. B., Ohte, N., & Boyer, E. W. (2008). Carbon isotope fractionation of dissolved inorganic carbon (DIC) due to outgassing of carbon dioxide from a headwater stream. *Hydrologic Processes*, *22*(14), 2410–2423. https://doi.org/10.1002/hyp.6833

Eimers, M. C., Watmough, S. A., & Buttle, J. M. (2008). Long-term trends in dissolved organic carbon concentration: A cautionary note. *Biogeochemistry*, *87*(1), 71–81. https://doi.org/10.1007/s10533-007-9168-1

Evans, C. D., & Monteith, D. T. (2001). Chemical trends at lakes and streams in the UK Acid Waters Monitoring Network, 1988–2000: Evidence for recent recovery at a national scale. *Hydrology and Earth System Sciences*, *5*(3), 351–366. https://doi.org/10.5194/hess-5-351-2001

Findlay, S. E. (2005). Increased carbon transport in the Hudson River: Unexpected consequence of nitrogen deposition? *Frontiers in Ecology and the Environment*, *3*, 133–137. https://doi.org/10.1890/1540-9295(2005)003[0133:ICTITH]2.0.CO;2

Finstad, A., Andersen, T., Larsen, S., Tominaga, K., Blumentrath, S., de Wit, H. A., et al. (2016). From greening to browning: Catchment vegetation development and reduced S-deposition promote organic carbon load on decadal time scales in Nordic lakes. *Scientific Reports*, *6*(1), 31944. https://doi.org/10.1038/srep31944

Freeman, C., Evans, C., Monteith, D., Reynolds, B., & Fenner, N. (2001). Export of organic carbon from peat soils. *Nature*, *412*(6849), 785. https://doi.org/10.1038/35090628

Gnann, S. J., Woods, R. A., & Howden, N. J. K. (2019). Is there a baseflow Budyko curve? *Water Resources Research*, *55*(4), 2838–2855. https://doi.org/10.1029/2018WR024464

Goldstein, B. A., Carlson, D., & Bhavsar, N. A. (2018). Subject matter knowledge in the age of big data and machine learning. *JAMA Network Open*, *1*(4):e181568. https://doi.org/10.1001/jamanetworkopen.2018.1568

Goldstein, B. A., & Rigdon, J. (2019). Using machine learning to identify heterogeneous effects in randomized clinical trials—Moving beyond the forest plot and into the forest. *JAMA Network Open*, *2*(3), e190004. https://doi.org/10.1001/jamanetworkopen.2019.0004

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., & Andréassian, V. (2014). Large-sample hydrology: A need to balance depth with breadth. *Hydrology and Earth System Sciences*, *18*(2), 463–477. https://doi.org/10.5194/hess-18-463-2014

Hanley, J. P., Rizzo, D. M., Buzas, J. S., & Eppstein, M. J. (2020). A tandem evolutionary algorithm for identifying causal rules from complex data. *Evolutionary Computation*, *28*(1), 87–114. https://doi.org/10.1162/evco_a_00252

Hanley, J. P., Rizzo, D. M., Stevens, L., Cahan, S. H., Dorn, P. L., Morrissey, L. A., et al. (2020). Novel evolutionary algorithm identifies interactions driving infestation of *Triatoma dimidiata*, a Chagas disease vector. *The American Journal of Tropical Medicine and Hygiene*, *103*(2), 735–744. https://doi.org/10.4269/ajtmh.18-0733. Retrieved from https://www.ajtmh.org/view/journals/tpmd/103/2/article-p735.xml

Hayhoe, K., Wake, C. P., Huntington, T. G., Luo, L., Schwartz, M., Sheffield, J., et al. (2007). Past and future changes in climate and hydrological indicators in the U.S. Northeast. *Climate Dynamics*, *28*(4), 381–407. https://doi.org/10.1007/s00382-006-0187-8

Herlihy, A. T., Paulson, S. G., Van Sickle, J., Stoddard, J. L., Hawkins, C. P., & Yuan, L. L. (2008). Striving for consistency in a national assessment: The challenges of applying a reference-condition approach at a continental scale. *Journal of the North American Benthological Society*, *27*(4), 860–877. https://doi.org/10.1899/08-081.1

Hirsch, R. M. (2011). A perspective on nonstationarity and water management. *Journal of the American Water Resources Association*, *47*(3), 436–446. https://doi.org/10.1111/j.1752-1688.2011.00539.x

Hruška, J., Krám, P., McDowell, W. H., & Oulehle, F. (2009). Increased dissolved organic carbon (DOC) in Central European streams is driven by reductions in ionic strength rather than climate change or decreasing acidity. *Environmental Science & Technology*, *43*(12), 4320–4326. https://doi.org/10.1021/es803645w

IPCC. (2013). *Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Jenks, G. F. (1967). The data model concept in statistical mapping. *International Yearbook of Cartography*, *7*, 186–190.

JMP. (2019). *Version 15*. SAS Institute Inc.

Knorr, K.-H. (2013). DOC-dynamics in a small headwater catchment as driven by redox fluctuations and hydrological flow paths – Are DOC exports mediated by iron reduction/oxidation cycles? *Biogeosciences*, *10*(2), 891–904. https://doi.org/10.5194/bg-10-891-2013

Larsen, L., Thomas, C., Eppinga, M., & Coulthard, T. (2014). Exploratory modeling: Extracting causality from complexity. *Eos, Transactions American Geophysical Union*, *95*(32), 285–292. https://doi.org/10.1002/2014EO320001

Lepistö, A., Kortelainen, P., & Mattsson, T. (2008). Increased organic C and N leaching in a northern boreal river basin in Finland. *Global Biogeochemical Cycles*, *22*(3), GB3029. https://doi.org/10.1029/2007GB003175

Lin, M., Lucas, H. C., & Shmueli, G. (2013). Research commentary —Too big to fail: Large samples and the *p*-value problem. *Information Systems Research*, *24*(4), 906–917. https://doi.org/10.1287/isre.2013.0480

Lins, H. F. (2012). *USGS Hydro-Climatic Data Network 2009 (HCDN-2009), US Geological Survey, Fact Sheet 2012-3047*. US Geological Survey.

MacDonald, L. H., & Coe, D. (2007). Influence of headwater streams on downstream reaches in forested areas. *Forest Science*, *53*(2), 148–168. https://doi.org/10.1093/forestscience/53.2.148

MATLAB. (2018). *9.7.0.1190202 (R2019b)*. The MathWorks Inc.

Monteith, D. T., Stoddard, J. L., Evans, C. D., De Wit, H. A., Forsius, M., Høgåsen, T., et al. (2007). Dissolved organic carbon trends resulting from changes in atmospheric deposition chemistry. *Nature*, *450*(7169), 537–540. https://doi.org/10.1038/nature06316

Murray, C. J. L., Vos, T., Lozano, R., Naghavi, M., Flaxman, A. D., Michaud, C., et al. (2012). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: A systematic analysis for the global burden of disease study 2010. *Lancet*, *380*(9859), 2197–2223. https://doi.org/10.1016/S0140-6736(12)61689-4

Musolff, A. (2020). WQQDB – Water quality and quantity data base Germany: Metadata. *HydroShare*. https://doi.org/10.4211/hs.a42addcbd59a466a9aa56472dfef8721

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydrometeorological dataset for the contiguous USA: Dataset characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. https://doi.org/10.5194/hess-19-209-2015

Omernik, J. M. (1987). Ecoregions of the conterminous United States. *Annals of the Association of American Geographers*, *77*(1), 118–125. https://doi.org/10.1111/j.1467-8306.1987.tb00149.x

Öquist, M. G., Bishop, K., Grelle, A., Klemedtsson, L., Köhler, S. J., Laudon, H., et al. (2014). The full annual carbon balance of boreal forests is highly sensitive to precipitation. *Environmental Science and Technology Letters*, *1*(7), 315–319. https://doi.org/10.1021/ez500169j

Öquist, M. G., Wallin, M. B., Seibert, J., Bishop, K., & Laudon, H. (2009). Dissolved inorganic carbon export across the soil/stream interface and its fate in a boreal headwater stream. *Environmental Science & Technology*, *43*(19), 7364–7369. https://doi.org/10.1021/es900416h

Pelletier, J. D., Broxton, P. D., Hazenberg, P., Zeng, X., Troch, P. A., Niu, G.-Y., et al. (2016). A gridded global data set of soil, intact regolith, and sedimentary deposit thicknesses for regional and global land surface modeling. *Journal of Advances in Modeling Earth Systems*, *8*(1), 41–65. https://doi.org/10.1002/2015MS000526

Perdrial, J. N., McIntosh, J., Harpold, A., Brooks, P. D., Zapata-Rios, X., Ray, J., et al. (2014). Stream water carbon controls in seasonally snow-covered mountain catchments: Impact of inter-annual variability of water fluxes, catchment aspect and seasonal processes. *Biogeochemistry*, *118*(1–3), 273–290. https://doi.org/10.1007/s10533-013-9929-y

Perdrial, J. N., Thompson, A. A., & Chorover, J. (2015). Soil geochemistry in the critical zone: Influence on atmosphere, surface- and groundwater composition. In J. R. Giardino & C. Houser (Eds.), *Principles and dynamics of the critical zone* (pp. 173–201). Elsevier.

Porcal, P., Koprivnjak, J. F., Molot, L. A., & Dillon, P. J. (2009). Humic substances—Part 7: The biogeochemistry of dissolved organic carbon and its interactions with climate change. *Environmental Science & Pollution Research*, *16*(6), 714–726. https://doi.org/10.1007/s11356-009-0176-7

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rabosky, D. L., Grundler, M. C., Anderson, C. J., Title, P. O., Shi, J. J., Brown, J. W., et al. (2014). BAMMtools: An R package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods in Ecology and Evolution*, *5*(7), 701–707. https://doi.org/10.1111/2041-210X.12199

Rasmussen, C., Matsuyama, N., Dahlgren, R. A., Southard, R. J., & Brauer, N. (2007). Soil genesis and mineral transformation across an environmental gradient on Andesitic Lahar. *Soil Science Society of America Journal*, *71*(1), 225–237. https://doi.org/10.2136/sssaj2006.0100

Raymond, P., Hartmann, J., Lauerwald, R., Sobek, S., McDonald, C., Hoover, M., et al. (2013). Global carbon dioxide emissions from inland waters. *Nature*, *503*(7476), 355–359. https://doi.org/10.1038/nature12760

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Schlesinger, W. H., & Melack, J. M. (1981). Transport of organic carbon in the world's rivers. *Tellus*, *33*(2), 172–187. https://doi.org/10.1111/j.2153-3490.1981.tb01742.x

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., et al. (2018). HESS opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences*, *22*(11), 5639–5656. https://doi.org/10.5194/hess-22-5639-2018

Steele, M. K., & Aitkenhead-Peterson, J. A. (2012). Urban soils of Texas: Relating irrigation sodicity to water-extractable carbon and nutrients. *Soil Science Society of America Journal*, *76*(3), 972–982. https://doi.org/10.2136/sssaj2011.0274

Sterle, G., Perdrial, J., Li, L., Adler, T., Underwood, K., Rizzo, D., et al. (2022). CAMELS-Chem: Augmenting CAMELS (Catchment Attributes and Meteorology for Large-sample Studies) with atmospheric and stream water chemistry data. Hydrology and Earth System Sciences Discussions [preprint]. https://doi.org/10.5194/hess-2022-81

Urbanowicz, R. J., Andrew, A. S., Karagas, M. R., & Moore, J. H. (2013). Role of genetic heterogeneity and epistasis in bladder cancer susceptibility and outcome: A learning classifier system approach. *Journal of the American Medical Informatics Association*, *20*(4), 603–612. https://doi.org/10.1136/amiajnl-2012-001574

U.S. EPA. (2015). National Wetland Condition Assessment 2011: A collaborative survey of the Nation's Wetlands. EPA-843-R-15-005. Retrieved from https://www.epa.gov/sites/production/files/2015-11/documents/nwca_2011_draft_public_report_oct2015_v1_0.pdf

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$". *The American Statistician*, *73*(Suppl. 1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Wen, H., Perdrial, J., Bernal, S., Abbott, B. W., Bernal, S., Dupas, R., et al. (2020). Temperature controls production but hydrology controls export of dissolved organic carbon at the catchment scale. *Hydrology and Earth System Sciences Discussions*, *24*(2), 945–966. https://doi.org/10.5194/hess-24-945-2020

Worrall, F., & Burt, T. P. (2007). Trends in DOC concentration in great Britain. *Journal of Hydrology*, *346*(3–4), 81–92. https://doi.org/10.1016/j.jhydrol.2007.08.021

Yallop, A. R., & Clutterbuck, B. (2009). Land management as a factor controlling dissolved organic carbon release from upland peat soils 1: Spatial variation in DOC productivity. *The Science of the Total Environment*, *407*(12), 3803–3813. https://doi.org/10.1016/j.scitotenv.2009.03.012