Classifying Annual Daily Hydrographs in Western North America using t-Distributed Stochastic Neighbor Embedding (t-SNE)

Weigang Tang^1 and $\operatorname{Sean} \operatorname{Carey}^1$

¹McMaster University

September 27, 2021

Abstract

Flow regimes are critical for determining physical and biological processes in rivers, and their classification and regionalization traditionally seeks to link patterns of flow to physiographic, climate and other information. There are many approaches to, and rationales for, catchment classification, with those focused on streamflow often seeking to relate a particular response characteristic to a physical property or climatic driver. Rationales include such topics as Prediction in Ungauged Basins (PUB), helping with experimental approaches, and providing guidance for model selection in poorly understood hydrological systems. While scale and time are important considerations for classification, the Annual Daily Hydrograph (ADH) is a first-order easily visualized integrated expression of catchment function, and over many years is a distinct hydrological signature. In this study, we use t-SNE, a state-of-the-art technique of dimensionality reduction, to classify 17110 ADHs for 304 reference catchments in mountainous Western North America. t-SNE is chosen over other conventional methods of dimensionality reduction (e.g. PCA) as it presents greater separability of ADHs, which are projected on a 2D map where the similarities are evaluated according to their map distance. We then utilize a Deep Learning encoder to upgrade the non-parametric t-SNE to a parametric approach, enhancing its capability to address 'unseen' samples. Results showed that t-SNE was an effective classifier as it successfully clustered ADHs of similar flow regimes on the 2D map. In addition, many compact clusters on the 2D map in the coastal Pacific Northwest suggest information redundancy in the local hydrometric network. The t-SNE map provides an intuitive way to visualize the similarity of high-dimensional data of ADHs, groups catchments with like characteristics, and avoids the reliance on subjective hydrometric indicators.

Introduction

Hydrologists typically acquire process knowledge from detailed place based studies and from representative experimental catchments, where hydrometric and biophysical attributes can be intensively measured over time. There are are large number of global catchment observation networks, yet in many parts of the world they are in decline due to the expense in establishing, operating and maintaining their infrastructure (Laudon et al., 2017). Consequently, extrapolating process knowledge to watersheds that are hydrologically similar, yet not necessarily measured, has been a major focus of the hydrological community for the past several decades with initiatives such as the Prediction in Ungauged Basins (PUB) program (Sivapalan et al., 2003), whose goal was to predict flow quantiles at ungauged or poorly gauged basins according to the historical flow data collected at hydrologically similar basins.

Catchment classification has a long history as a means to generalize the functional behaviour that exists within watersheds, quantify their similarity, and to transfer information among them (Wagener et al., 2007). While there is no universal hydrological classification, the degree of similarity that exists is often defined from intrinsic and response characteristics of watersheds such as: climate (e.g. temperature, precipitation), watershed biophysical characteristics (e.g. geological conditions, soil type, relief, and vegetation), and the

flow regime (e.g. annual hydrograph). Climate indices for classification (e.g. K["]oppen, Thornthwaite) are widely applied at varying time scales and have an extremely long history identifying the intrinsic seasonality, thermal and moisture regimes of a region. Physiographic and biophysical indices such as soils, topography and geology strongly influence catchment behaviour (Buttle, 2006; Bormann, 2010), yet are not always ideal in defining process controls on catchment behaviour across scales and regions (Merz and Bloschl, 2005). Often, catchments with similar climate and physical conditions are not hydrologically similar (Oudin et al., 2010; Ali et al., 2012).

Evaluating catchment similarity based solely in terms of streamflow characteristics is popular; particularly in aquatic ecology where habitats are particularly sensitive to flow regimes (Poff et al., 1997). However, as Sawicz et al. (2011) notes, ecological studies are not typically aimed at understanding the behaviour of the catchment including the causes of a particular regime. Over time, the flow regime of a catchment is a descriptor of the seasonal behaviour of the streamflow (Haines et al., 1988) and by its nature is an integrator of a variety of hydrological processes produced by the interaction between climate and catchment physical characteristics. After decades of development, there are hundreds of indices available which quantitatively characterize five major components of flow regime: magnitude, timing, duration, frequency, and rate of change (Poff et al., 1997). Flow statistics (e.g. mean, max, and quantiles, standard deviation) at varying temporal scale are widely-used indices that reveal first-order information regarding magnitude, distribution, and variation of stream flow over a period of interest (Hall and Minns, 1999; Carey et al., 2010; Ali et al., 2012; Toth, 2013). More sophisticated indices, often explicitly reflecting specific hydrological processes, are preferred in catchment classification with respect to hydrological functions and system complexity (Sawicz et al., 2011). However, it remains a challenge to design a combination of hydrological indices that fully describe dominant hydrological characteristics of flow regimes, maximize distinctiveness among different flow regimes, as well as avoid information redundancy.

Classification based on flow statistics using clustering algorithms such as C-means and artificial neural networks (ANN) (Hall and Minns, 1999), hierarchical models (Snelder et al., 2005), and Bayesian clustering algorithm (Kennard et al., 2010; Sawicz et al., 2014), have been successfully applied for catchment classification and regionalization. The premise is to identify groups (or regions) in a way that similarity within a region is maximized whereas similarity between regions is minimized. Self-organized mapping (SOM), an unsupervised ANN machine learning technique has become increasingly appealing as it produces a low dimensional (typically two) representation of higher dimensional data that is simple to visualize. SOM preserves the topological structure of data as it transforms information from high-dimension feature space, and clusters information visually on maps where clustered points are more similar that distal points. When hydrological indices are transformed, catchments with homogeneous features are close on the 2-D map, and distance on the map can be used to visually infer similarity (Di Prinzio et al., 2011; Ley et al., 2011; Razavi and Coulibaly, 2013; Toth, 2013). Previously, SOM has been applied for catchment grouping with a moderate (50) number of samples (Ley et al., 2011; Toth, 2013), yet for extremely large data sets with thousands or millions of samples, computational time increases with sample size, challenging the utility of SOM application for very large data sets.

The objective of this research is to design and implement a novel method to visualize and classify streamflow regimes for a large streamflow data set focused on undisturbed rivers western North America. The classification is based on annual daily hydrographs (ADHs) from 304 sites over multiple years, providing 17110 ADHs for classification. The large nature of this data set renders traditional SOM impractical, and we therefore utilize t-distribution Stochastic Neighbor Embedding (t-SNE), an alternative machine learning algorithm proposed by van der Maaten (2009), to map ADHs to 2D feature space to assess flow similarity and compare this to traditional Principal Component Analyses. Furthermore, we develop an encoder neural network that allows additional data to be projected on to the t-SNE map; overcoming previous challenges with the non-parametric t-SNE technique. While this methodology only focuses on a limited region and does not attempt a universal classification, we attempt to show the novelty, flexibility and potential of this approach for future classification activities.

Methods

Data

Daily flow data were gathered from Western North America (WNA) stream gauges in the Referenced Hydrometric Basin Network (RHBN) and Hydro-climate Data Network (HCDN), which are operated by Water Survey of Canada (WSC) and United States Geological Survey (USGS), respectively. Streams selected for the networks have predominantly natural flow regimes, with minimum human disturbance (e.g. significant land-use change, dams, reservoirs, and hydro-power stations) during long-term observation periods. In total, 304 HCDN and RHBN stream gauges were deemed suitable in WNA, including four Canadian provinces/territories (British Columbia, Alberta, Yukon, and Northwest Territory) and eleven American states (Washington, Oregon, California, Idaho, Nevada, Montana, Utah, Wyoming, Colorado, Arizona and New Mexico) (Fig. 1). The observation length varies among gauges, ranging from ten to over one hundred years. At each site, the long-term daily hydrograph was broken into the Annual Daily Hydrographs (ADHs). Each ADH contains 365 values of daily flow over a year from 1 January to 31 December (leap days excluded if applicable). Small gaps ([?]7 days) in ADHs were filled up via linear interpolation. ADHs with gaps > 7 consecutive days were excluded.

Figure 1: Selected HCDN and RHBN streamflow gauges in western North America

During initial data screening, ADHs with atypical shapes (e.g. sudden zig-zag patterns) were detected and excluded due to uncertainty in data quality. Furthermore, we excluded the ADHs of extreme years as they are less representative to the general flow pattern. In this case, extreme wet years are defined as ADHs with a maximum flow one hundred times higher than its long-term average, while extreme dry years are ADHs with a minimum flow 100 times smaller than its long-term average. From an initial set of 19499 ADHs, 17110 ADHs were preserved for analysis.

Selected ADHs were pre-processed prior to analysis. First, ADHs for a given stream were divided by its longterm average, which helps limit the scaling factor of watershed size and enhance comparability. Subsequently, a log transformation was applied to reduce skewness of the data as machine learning algorithms typically have improved performance on normally distributed data. A small number (10^{-6}) was added to ADHs to avoid invalid values during log transformation (see Eq. 1). Furthermore, min-max normalization (see Eq. 2) was applied to scale values of ADHs into a range of 0 and 1, which was recommended by the original paper of t-SNE (van der Maaten and Hinton, 2008).

$$f(x) = log(x + e^{-6}) (1$$

$$f(x) = \frac{x - x_{min}}{x_{max} - x_{min}}$$

(2)

Labelled Samples

A subset of ADHs were manually labelled with flow regime types. The determination of flow regime type was based on visual assessment to the shape of ADHs, which is subjective to some degree. 2480 ADHs were labelled that represent seven distinctive flow regimes in the WNA domain (Fig. 2). ADHs selected for the same type typically come from streams that are geographically clustered. Class 1 is characterized by frequent rain events during winter and low flows in summer, with streams mostly located in coastal Pacific North

West (PNW). Class 2 is similar to Class 1, but has a distinct snowmelt-driven spring freshet, and are mainly located in the interior PNW. Class 3 has extremely low winter flows, large spring freshet and many summer rain events. All of Class 3 gauges are located in Alaska. Class 4 exhibits a large spring freshet followed by summer storm events with gauges located in the Northwest Territories and northern Alberta. Class 5 is similar to Class 4, with later summer rain events and gauges are located in Yukon and northern British Columbia. Class 6 is characterized by dominant snowmelt freshet in spring that accounts for more than 70% of the annual discharge with gauges primarily located Canadian Rockies between British Columbia and Alberta. Class 7 has a spring freshet along with occasional winter events and gauges are most often located in Idaho. The labelled ADHs were used to evaluate the performance of t-SNE, with details provided in later sections.

Figure 2: Normalized ADHs for the seven flow regimes. The number in the parenthesis indicates the number of ADHs in that class.

t-distribution Stochastic Neighbor Embedding (t-SNE)

t-distribution Stochastic Neighbor Embedding (t-SNE, van der Maaten and Hinton (2008)) is a state-oftheart technique for dimensionality reduction and high-dimensional data visualization. It is a variant of the SNE that was originally proposed by Hinton and Roweis (2002). SNE represents similarity between datapoints using conditional probabilities that are converted from pairwise Euclidean distances (Eq. 3).

$$p_{i|j} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$

(3)

where x_i and x_j denote datapoints in the original feature space, $kx_i - x_j k^2$ is the Euclidean distance, p_{i+j}

is the conditional probability between x_i and x_j , and σ_i is the variance of the Gaussian distribution centered at x_i . The value of σ_i primarily depend on the data density at x_i , which varies among datapoints. Details on how to determine the value of σ_i can be found in van der Maaten and Hinton (2008). The conditional probabilities between the low-dimensional counterpart y_i and y_j is indicated as $q_i \mid_j$. Through minimizing the differences between $p_i \mid_j$ and $q_i \mid_j$, SNE maximally copies the local and global structure of datapoints from orginal to new feature space. Kullback-Leibler Divergence (KLD), a measure of difference between probabilities, is employed by SNE as the loss function (see Eq. 4).

$$KL(P \parallel Q) = \sum_{i} \sum_{j} p_{ij} \cdot \log \frac{p_{ij}}{q_{ij}}$$

(4)

Compared with original SNE, t-SNE has two major adjustments: 1) the use of joint probabilities (Eq. 5) instead of conditional probabilities to represent similarities, and 2) the use of a Student t-distribution instead of Gaussian distribution to compute similarities between two datapoints in low-dimensional space (Eq. 6).

 $p_{i|j}$ $p_{j|i}$ p_{ij}

(5)

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$$

(6)

where N is the total number of datapoints, and p_{ij} and q_{ij} denote the joint probability in high- and lowdimensional space. The first adjustment simplifies the form of loss function improving training efficiency. The second adjustment effectively alleviated the "crowding problem", a common challenge faced by SNE and many other dimensionality reduction techniques (Goodfellow et al., 2016). Datapoints that lie moderately distant in high-dimensional space tend to be crushed together in the embedded space, which prevents gaps from forming between natural clusters. Using a heavy-tailed distribution (i.e. Student t-distribution) to calculate pairwise similarities for low-dimensional representation effectively alleviates the crowding problem and preserves the local and global structure of datapoints in the embedded space. In all trials of van der Maaten and Hinton (2008), t-SNE produced considerably better visualization than other embedding techniques, including SNE, Sammon mapping, curvilinear components analysis, Isomap, maximum variance unfolding, locally linear embedding, and Laplacian Eigenmaps.

Datapoint arrangement on t-SNE maps are sensitive to two parameters: 1) perplexity and 2) learning rate. The perplexity is considered a smooth measure of the effective number of neighbors and is used to determine σ_i of the Gaussian distribution for high-dimensional datapoints (van der Maaten and Hinton, 2008). Intuitively speaking, it controls the "roundness" of the arrangement of the datapoints in the embedded space. Conventionally, perplexity is chosen from a range between 5 and 50, with van der Maaten and Hinton (2008) recommending a perplexity of 30. Learning rate (η) is an optimization that determines the final convergence of loss function. In this study, we follow the process of van der Maaten and Hinton (2008) and choose perplexity as 30 and η as 100.

Results from the t-SNE map were compared with Principle Component Analysis (PCA), likely the most popular linear embedding technique and evaluated based on the separability of ADHs with distinct flow patterns. A subset of ADHs were labelled with flow regime type as detailed earlier, and then projected on the 2D map using t-SNE and PCA respectively. K-Nearest Neighbors (KNN, Goldberger et al., 2005) was used to classify datapoints on both embedded maps, with k set to 30. Favorable embedding techniques should arrange the ADHs of different types into separable clusters on 2D map and allow accurate classification with simple classifiers (e.g. KNN). Here, we employed classification accuracy as a quantitative indicator for separability of datapoints, and better embedding technique should result in higher accuracy. Tools for t-SNE, PCA, and KNN are all available in scikit-learn package in Python (Pedregosa et al., 2011).

As a non-parametric method, t-SNE does not produce the mapping function between high- and lowdimensional data representation. It is impossible to project additional data on to an existing t-SNE map, which has been recognized as a major shortcoming of t-SNE technique (van der Maaten and Hinton, 2008). One solution is to merge new data to the original dataset and re-run t-SNE, yet it is computationally inefficient for large datasets. In addition, datapoints in the original dataset could be displaced on the new t-SNE map as inclusion of new data alters the similarity matrix (i.e. pairwise joint probabilities). The location inconsistency of embedded datapoint is highly unfavourable. To address this, van der Maaten (2009) proposed an alternative solution that build an parametric t-SNE through incorporating an autoencoder. sub

Autoencoder

An autoencoder is a neural network that is trained to copy inputs to outputs through the use of an encoder and decoder. The encoder function $\mathbf{h} = f(\mathbf{x})$ converts input data \mathbf{x} to latent features \mathbf{h} , while the decoder function $\mathbf{r} = g(\mathbf{h})$ reconstruct data from latent features to its original format. As a lossy technique, autoencoders are not trained to copy perfectly, but to transfer the most salient information of input data to latent features (with less dimensions) while ignoring noise. This approach is widespread in dimensional reduction and feature learning (Goodfellow et al., 2016).

Encoder networks are often considered a universal function approximators. They are trained to approximate complicated, non-linear functions that map high-dimensional data to low-dimensional representation. Here, we use an encoder network to approximate the mapping between ADHs and t-SNE 2D data. Consequently, newly collected ADHs can be projected on the existing t-SNE map using the trained encoder, and there is no need to re-run t-SNE with the entire dataset. The objective is to minimize the distances between the datapoints projected by the t-SNE and encoder. Mean Absolute Error (MAE) was employed as the loss function.

The performance of the encoder network largely depends on the choice of model hyperparameters, which are referred to as untrainable parameters as they do not change during the training procedure. The network architecture (i.e. number of layers and nodes) and activation function are critical hyperparameters for the encoder. In an iterative manner, we tested a variety of architectures and activation functions in order to search for the optimal model configuration (see Table 2). Network architecture defines the depth (i.e. number of layers) and width (i.e. number of nodes) of the encoder network. Our baseline architecture is a three-layer fully-connected network, with 512, 256, and 128 nodes for each layer. Activation functions bring non-linearity to neural networks, and enable the encoder model to approximate complicated, non-linear functions. The choice of activation function affects the optimization process (Goodfellow et al., 2016), and we selected the Rectified Linear Unit (ReLU, as $f(x) = \max(0, x)$) as the default function for the baseline model Goodfellow et al. (2016). A modified version of ReLU, called leaky ReLU (as $f(x) = \max(a^* x, x)$), has been recommended as providing additional network optimization benefits (?), and a number were tested with tuned a ranging from 0.02 to 0.4. A dropout layer (Srivastava et al., 2014) was added after dense layers to avoid overfitting. The optimizer was chosen as Adam (Kingma and Ba, 2015) for all tested encoders. The encoder networks were built using Keras (2.4.3 version), a Python package for deep learning (Chollet et al., 2015).

Results

t-SNE vs PCA

Following the normalization of the ADHs, a PCA and t-SNE were implemented on the labelled data (Fig. 3). The PCA suffered from the "crowding issue", with points clustered within a relatively narrow region on the 2D map. In contrast, t-SNE had points that were more evenly dispersed across the 2D map, yet with distinct clusters. Classes 1, 2 and 7 were notably separate on both PCA and t-SNE maps, with class 6 more separable on the t-SNE map. Classes 3-5 were mixed together on both maps, yet they form an individual cluster on the t-SNE map that is isolated from the other clusters. All three of these classes were dominated

by gauges in the northern part of the domain (Alaska, Yukon, Northwest Territories, and northern British Columbia), highlighting their similarity (Fig. 2). Consequently, we merged Classes 3-5 into one (as Class 5) for subsequent analyses.

Figure 3: 2D map of PCA and t-SNE with labelled ADHs of seven flow regimes. Color of data points indicates flow regime type.

Based on the ADH arrangement on 2D map and KNN classification accuracy, t-SNE was superior to PCA. The PCA and t-SNE transformed data were classified using KNN, with a classification accuracy of 84.5% (PCA) and 90.3%(t-SNE), indicating the t-SNE map had enhanced separability among the flow regimes. A confusion matrix (Tab. 1) indicated that misclassification most often occurred among ADHs from Class 2 (the interior PNW) and Class 7 (predominantly Idaho), which are geographically proximal. This is reasonable as both flow regimes were characterized by a large spring freshet and winter storm events, yet flows in Class 7 typically had a more dominant freshet and less winter events than Class 2. In certain years dominated by either greater/less snow and rain in the winter resulted in ADHs resembling the other class. This highlights the challenge of subjectively labelling ADHs for classification, and suggests that refining the selection of labelled AHDs for these two classes may be warranted.

Figure 4: KNN-classified data points on PCA and t-SNE map. Black points represent misclassified ADHs.

| | Class 1 | Class 2 | Class 5 | Class 6 | Class 7 |
|---------|---------|---------|---------|---------|---------|
| Class 1 | 0.996 | 0.004 | 0.000 | 0.000 | 0.000 |
| Class 2 | 0.043 | 0.823 | 0.000 | 0.003 | 0.131 |
| Class 5 | 0.000 | 0.000 | 0.989 | 0.011 | 0.000 |
| Class 6 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 |
| Class 7 | 0.000 | 0.019 | 0.000 | 0.000 | 0.981 |

Table 1: Confusion matrix of KNN classification with t-SNE datapoints

Global t-SNE Map

A global t-SNE map was constructed with all ADHs, with labelled ADHs highlighted in color based on their flow regime (Fig. 5). ADHs from the same flow regime remained clustered on the map, yet the absolute location of labelled ADHs has changed from the previous map (Fig. 3). However, the topological relations among the clusters of flow regimes were largely preserved. For instance, the clusters for Class 1 and 2, which have similar ADH shape, remained close on the global t-SNE map, while those that were distinct (e.g. Class 1 and 6) are widely spaced across the map.

Figure 5: t-SNE map with all ADH samples. ADHs with flow regime labels are highlight with colors.

A large number of compact clusters on the t-SNE map were observed for Classes 1 and 2 as the ADHs in these clusters were highly correlated (Fig. 6). Approximately 50 of those clusters were scrutinized, and they result from same-year ADHs from a number of geographically proximate watersheds located near the Pacific coast (i.e. southern BC, Washington, Oregon, and northern California). This suggests that climate for a given year drives similar hydrological responses across these watersheds, which have highly similar responses.

There are 24 streams whose ADHs consistently occur in these clusters.

Figure 6: Small, compact clusters of ADHs are mostly observed in the division of Class 1 and 2 (a). ADHs of the three selected clusters exhibit strong correlation within clusters (b, c, d). Streams whose ADHs are frequently included in the compact clusters are all located in coastal PNW region (e).

Encoder

A number of encoders were tested, and the encoder with best MAE on the testing dataset employed a ninelayer architecture with activation functions of LeakyReLU (alpha=0.07) (see Encoder 50 in Table 2). The Dropout layer slightly improved loss and remarkably reduced the gap between the training and validating sets. This trained encoder achieved a MAE of 1.07 on training set, and 3.81 on testing set. The average displacement between the projected ADH points and their encoded counterparts on the t-SNE map was 5.93 for the testing set, which was small compared to the extent of the map (Fig. 7). The points formed a near-circular shape distribution on t-SNE map (see Fig. 5), so the map extent was measured by a radius to the map centroid. The radius that covers 98% data points on t-SNE map was 115.9. The ratio of average displacement to map radius is only 0.05, suggesting the errors of encoder projection are limited. After eliminating outliers (i.e.> 95 percentile), the average displacement for testing set dropped to 4.36.

Figure 7: Lines on map indicate the displacement between original t-SNE points and their counterpart projected by encoder. The displacements for training set are very small, while those for testing set are relatively large.

To illustrate the potential use of an encoder and the t-SNE map for classification, an ADH with an unknown flow regime was randomly selected and projected onto the t-SNE map using the optimal encoder (Fig. 8). Its ten nearest neighbours were identified and ADHs plotted with the unknown ADH (Fig. 8). Based on this procedure, the ADH is presumed to belong in Class 6; a flow regime that resembles the snow-dominated regime of the Canadian Rockies. The implications of this procedure are discussed below.

Figure 8: A randomly selected ADH on t-SNE map projected by Encoder (a). Original (b) and normalized ADHs (c) of ten nearest neighbors.

Discussion

The t-SNE map provides an novel and intuitive way to visualize similarities for large sets of stream hydrographs, as ADHs with similar shapes remain close on the 2D map. Here, we propose that the distance between points on the 2D map can be used as a similarity metric among watersheds, or for a given watershed the space it occupies on the map can be used to infer its relative variability in flows. In terms of grouping watersheds, t-SNE is superior to PCA with respect to separability of ADHs from different flow regimes. Furthermore, t-SNE is particularly suited to large data sets, requires less computational power, and is more interpretable than conventional visualization tools (i.e. pairwise similarity matrix). While we have only used data from western North America, t-SNE can be more broadly applied to larger or more constrained data sets.

If new hydrographs are obtained, ADHs can be projected on the t-SNE map with the trained encoder, allowing hydrographs to be quickly associated with like counterparts and information, such as flow regime type, seasonal pattern, and dominant hydrological processes can be estimated according to its nearest neighbors on t-SNE map. In this way, this approach can be used for comparative analysis, flow regime classification and regionalization and potentially for change detection.

A Novel Similarity Metric

To confirm the validity of t-SNE distance as a metric of similarity, the pairwise t-SNE distances between ADHs on the 2-D map are compared with the counterparts of cross correlations (Xcorr), which is a conventional metric used to measure similarity between time series. Xcorr is calculated for every pair of ADHs. For every single ADH, the average of Xcorr for all other ADHs is 0.20, while that for its 10 nearest neighbor is 0.87. It manifest that ADHs close on t-SNE map indeed share considerable similarity of flow regime pattern. Regression analysis between t-SNE distance and Xcorr indicates a significant relationship (p-value < 0.01) with a Spearman R of -0.80. Nash-Sutcliffe Efficiency (NSE) is another widely-used metric that measure the

consistency between hydrographs. The relationship between t-SNE distance and NSE is also statistically significant and with Spearman R of -0.75.

Cross Correlation Matrix (CCM) is an alternative method to visualize pairwise similarity between hydrographs, yet is impractical for large datasets. Our dataset produces a 17110-by-17110 CCM, with more than 140 million entries (after removing twins), making it computationally expensive to generate and difficult to recognize and interpret patterns at this scale. On the other hand, the t-SNE map provides an intuitive and efficient way to visualize similarity of ADHs, which is simply indicated by their distance on the map.

Misclassification of Mixed Regimes

Misclassification is a ubiquitous feature of machine learning algorithms, and Class 2 had the highest misclassification rate in our procedure. This is unsurprising as Class 2 is a mixed regime with both snowmelt driven freshet and high flows in winter, is a relative superposition of Class 7 and Class 1, and lies between them on the 2-D map. Depending upon winter temperatures, the ADHs in Class 2 tend to shift towards Class 1 in warm years and Class 7 in cold years, which is expected as warm anomalies bring rain (which prevails in the PNW) and cold anomalies enhance snow accumulation and melt (such as in the Canadian Rockies). We presume that watersheds that are more sensitive to climate anomalies are also more subject to misclassification.

Labeling Strategy

The performance of the t-SNE map is based on the separability of ADHs among different flow regimes, and the accuracy of the KNN classification provides a quantitative measure of this separability. Data quality of labelled ADHs is critical as it directly links to classification accuracy. While our strategy allowed a large number of samples to be labelled in a practical time frame, there is the potential for mislabeling ADHs for atypical years. The general nature of hydrological variability can impart a large variance on ADHs, and that in some years hydrographs have flow patterns that are unrepresentative of their labelled class. This inherent variability also explains why some samples from each class are apart from the majority on the t-SNE map.

Manually labelling ADHs is time consuming and there is considerable subjectivity and process knowledge that is required for meaningful classification. To remove subjectivity, Generative Adversarial Networks (GAN) have been used to artificially generate samples that closely mimic real ones (Goodfellow et al., 2014). GAN samples preserve the main patterns of the training data and introduce some random variation, and have considerable potential to create high-quality labelled samples of ADHs. Fed with a limited number of ADHs, GAN can generate an infinite number of samples for each flow regime, enlarging the sample size and reducing the influence of human subjectivity. While not used in this work, we suggest GAN is a promising tool for improving classification.

Optimal Encoder Selection

An effective encoder is a critical component of this approach as it determines the reliability of the mapping function between ADHs and the t-SNE data points and allows insertion of new data onto the t-SNE map. In this work, we tested 55 encoder models with various network architectures and activation functions (see Table 2) before selecting the optimal encoder (i.e. with minimum MAE) for ADH dataset. Here, we demonstrate the sensitivity of encoders to network depth and activation function type.

In many deep learning applications, deeper networks prevail due to their strong capability in recognizing and processing complicated patterns of data (Goodfellow et al., 2016). However, in our case, the encoder's performance did not consistently improve with the depth of network. A clear reduction in MAE was observed when increasing the number of layers from one to four, but further deepening of the layers did not show improvement (Fig. 9). The relative few number of layers was likely due to the relative simplicity of ADHs compared with photographic images, which are the most common application of DL research. Another method to enhance encoder performance was to widen the layers of the network, as increasing the nodes of the initial layer from 128 to 1024 provided marked improvement in MAE (Fig. 9). However, further expansion to 2048 nodes only marginally improved encoder's performance, suggesting wider layers are unnecessary. In the trials, we found the nine-layer network with an initial layer of 2048 nodes (N2048-L9, see Encoder 39 in Table 2) produced the lowest MAE of testing set, so this architecture is employed by the final encoder.

Using LeakyReLU as the activation function instead of ordinary ReLU substantially improved the encoder's performance. Decreasing the values of a from 0.4 to 0.02, we found the MAE of testing set continuously declined and reached to the minimum at a = 0.07, but with a slight increase below that (Fig. 9). With an appropriate a, LeakyReLU reduced the MAE by 10%. Our final encoder employed a LeakyReLU with a = 0.07 as the activation function.

Figure 9: Changes of encoders performance with width and depth of network (left) and alpha of LeakyReLU activation function (right).

Gaps of loss between the training and testing set are very common in supervised learning, as models always tend to perform less reliable on "unseen" data. However, huge gaps are often considered a sign of overfitting. In this case, there is a relatively large gap between the MAE of training and testing set, so Dropout algorithm (Srivastava et al., 2014) was applied for the top layers in selected the network in order to improve model generalization. During training, Dropout deactivate a number of randomly selected nodes (neurons) at each iteration. As a result, it prevents models from depending exclusively on certain neurons and therefore improve generalization. In this case, we found that Dropout effectively narrowed the gap of MAE between training and testing set. However, the reduction is mostly led by increasing loss of training set, while the loss of testing set largely remain the same. This indicated that model generalization had little improvement. Presumably, a more effective way to enhanced encoder generalization is to increase diversity and sample size of ADHs datasets.

Potential use of t-SNE maps

As a clustering technique, t-SNE is a powerful technique that has had very limited application to hydrological data sets (Mazher, 2020; Liu et al., 2021), yet there are many potential uses as 'big data' and machine learning emerge in hydrology. First, the t-SNE map provides an intuitive way to visualize the similarity of ADHs and other large time series data sets, and the encoder improves its practical use by turning the t-SNE from a non-parametric to a parametric method. It overcomes a major challenge of t-SNE by allowing new data to be projected on an existing map as it becomes available. The approach of information visualization can benefit a broad range of research that seeks to establish similarity among time series such and need not be restricted to streamflow hydrographs. For example, watershed classification, regionalization and streamflow change detection are all potential future areas of research using this methodology. t-SNE can also be used to infer similarity in other time series and signals to assess natural groupings and patterns.

Perhaps the most obvious use of t-SNE is in watershed classification as the parametric t-SNE technique can effectively identify natural groupings of watersheds and place new data within these groupings. While there is considerable research over the past few decades on hydrological classification (Wagener et al., 2007), t-SNE is a novel data-driven approach that allows fast identification of homogeneous and similar regimes, and process inference can be quickly transferred to their nearest neighbours. If the location is close enough on the t-SNE map, it would be possible to reconstruct hydrographs of ungauged watersheds according to their nearest neighbours (Patil and Stieglitz, 2012). Another potential utilization of t-SNE is identifying information redundancy, which is important in designing hydrometric and other monitoring networks (Coulibaly et al., 2013). For example, a number of highly homogeneous watersheds on the t-SNE map from the PNW suggest high consistency in the hydrographs and potential information redundancy in the network. However, the validity of information redundancy detected by t-SNE should be further verified with entropy-based approaches (Singh, 1997; Mishra and Coulibaly, 2009).

Conclusion

Here, we demonstrate the potential of t-SNE as an approach to compare similarity among hydrographs that is particularly useful for large data sets. By converting annual daily hydrographs to 2-D representation, their degree of similarity is indicated by distance on the map. t-SNE distance can be used as novel similarity metric, supplementing other comparative metrics such as Xcorr and Nash-Sutcliffe efficiency. In this application on reference watersheds in western North America, t-SNE outperformed PCA analysis for dimensional reduction, suggesting its potential in classifying and regionalizing streamflows. A deep learning encoder network was developed and trained to project new data onto existing maps to identify hydrological counterparts, overcoming previous challenges of the non-parametric t-SNE approach. While subjectivity in classification limits deep learning algorithms in some circumstances, we believe that t-SNE is underutilized in hydrological applications and that it has considerable potential for extremely large data sets.

Appendix I

Table 2: Model configuration of tested encoder

Acknowledgements

This research was financially supported by Global Water Futures program and the Natural Sciences and Engineering Research Council of Canada. The authors would like to thank the Water Survey of Canada and USGS for collecting, managing, and providing high-quality streamflow data across Canada and USA.

References

Ali, G., Tetzlaff, D., Soulsby, C., Mcdonnell, J. J., and Capell, R. (2012). A comparison of similarity indices for catchment classification using a cross-regional dataset. *Advances in Water Resources*, 40:11–22. Bormann, H. (2010). Towards a hydrologically motivated soil texture classification. *Geoderma*, 157(3-4):142–153. Buttle, J. (2006). Mapping first-order controls on streamflow from drainage basins: the t3 template. *Hydrological Processes*, 20(15):3415–3422. Carey, S. K., Tetzlaff, D., Seibert, J., Soulsby, C., Buttle, J., Laudon, H., ..., and Pomeroy, J. W. (2010). Inter-comparison of hydro-climatic regimes across northern catchments: synchronicity, resistance and resilience. *Hydrological Processes*, 24:3591–3602.

Chollet, F. et al. (2015). Keras. https://keras.io.

Coulibaly, P., Samuel, J., Pietroniro, A., and Harvey, D. (2013). Evaluation of canadian national hydrometric network density based on wmo 2008 standards. *Canadian Water Resources Journal*, 38(2):259–167. Di Prinzio, M., Castellarin, A., and Toth, E. (2011). Data-driven catchment classification: application to the pub problem.*Hydrology and Earth System Sciences*, 15:1921–1935.

Goldberger, J., Roweis, S., Hinton, G., and Salakhutdinov, R. (2005). Neighbourhood components analysis.

Advances in Neural Information Processing Systems, 17:513–520. Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning. MIT Press. http://www. deeplearningbook.org. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems, 2:2672–2680. Haines, A. T., Finlayson, B. L., and McMahon, T. A. (1988). A global classification of river regimes. Applied Geography, 8:255–272. Hall, M. J. and Minns, A. W. (1999). The classification of hydrologically homogeneous regions. Hydrological Sciences Journal, 44(5):693–704. Hinton, G. E. and Roweis, S. T. (2002). Stochastic neighbor embedding. In Advances in Neural Information Processing Systems, pages 833–840. MIT Press, Cambridge, MA, USA. Kennard, M. J., Pusey, B. J., Olden, J. D., Mackay, S. J., Stein, J. L., and Marsh, N. (2010). Classification of natural flow regimes in australia to support environmental flow management. Freshwater Biology, 55:171–193. Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations.

Laudon, H., Spence, C., Buttle, J., Carey, S. K., McDonnell, J. J., McNamara, J. P., Soulsby, C., and Tetzlaff, D. (2017). Save northern high-latitude catchments. Nature Geosciences, 10(5):324–325. Ley, R., Casper, M. C., Hellebrand, H., and Merz, R. (2011). Catchment classification by runoff behaviours with self-organizing maps (som). Hydrology and Earth System Sciences, 15:2947–2962. Liu, H., Yang, J., Ye, M., James, S. C., Tang, Z., Dong, J., and Xing, T. (2021). Using t-distributed stochastic neighbor embedding (t-sne) for cluster analysis and spatial zone delineation of groundwater geochemistry data. Journal of Hydrology, 597:126146. Mazher, A. (2020). Visualization framework for high-dimensional spatio-temporal hydrological gridded datasets using machine-learning techniques. Water, 12:590. Merz, R. and Bloschl, G. (2005). Flood frequency regionalisation - spatial proximity vs catchment attributes. Journal of Hydrology, 302(1-4):283-306. Mishra, A. K. and Coulibaly, P. (2009). Developments in hydrological network design: a review. Reviews of Geophysics, 47:RG2001. Oudin, L., Kay, A., Andreassian, V., and Perrin, C. (2010). Are seemingly physically similar catchments truly hydrologically similar. Water Resources Research, 46:W11558. Patil, S. and Stieglitz, M. (2012). Controls on hydrologic similarity: role of nearby gauged catchments for prediction at an ungauged catchment. Hydrology and Earth System Sciences, 16:551–562. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ..., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12:2825–2830. Poff, N. L., Allan, J. D., Bain, M. B., Karr, J. R., Prestegaard, K. L., Richter, B. D., Sparks, R. E., and Stromberg, J. C. (1997). The natural flow regime. *BioScience*, 47(11):769–784. Razavi, T. and Coulibaly, P. (2013). Classification of ontario watersheds based on physical attributes and streamflow series. Journal of Hydrology, 493:81–94. Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., and Carrillo, G. (2011). Catchment classification: emperical analysis of hydrological similarity based on catchment function in the eastern usa. Hydrological and Earth System Sciences, 15:2895–2911. Sawicz, K. A., Kelleher, C., Wagener, T., Troch, P., Sivapalan, M., and Carrillo, G. (2014). Characterizing hydrological change through catchment classification. Hydrology and Earth System Sciences, 18:273–285.

Singh, V. P. (1997). The use of entropy in hydrology and water resources. *Hydrological Processes*, 11:587–626.

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., ..., and Zehe, E. (2003). Iahs decade on predictions in ungauged basins (pub), 2003–2012: shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal*, 48(6):857–880. Snelder, T. H., Biggs, B. J. F., and Woods, R. A. (2005). Improved eco-hydrological classification of rivers. *River Research and Applications*, 21:609–628.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, L., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Toth, E. (2013). Catchment classification based on characterization of streamflow and precipitation time series. *Hydrology and Earth System Sciences*, 17:1149–1159. van der Maaten, L. (2009). Learning a parametric embedding by preserving local structure. *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, 5:384–391. van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605. Wagener, T., Sivapalan, M., Troch, P., and Woods, R. (2007). Catchment classification and hydrologic similarity. *Geography Compass*, 1/4:901–931.















Hosted file

appendix.xlsx available at https://authorea.com/users/438101/articles/539370-classifyingannual-daily-hydrographs-in-western-north-america-using-t-distributed-stochasticneighbor-embedding-t-sne