water

MDPI

Article

Visualization Framework for High-Dimensional Spatio-Temporal Hydrological Gridded Datasets using Machine-Learning Techniques

Abeer Mazher

Deep Earth Imaging—Future Science Platform (DEI-FSP), Commonwealth Scientific and Industrial Research Organization (CSIRO), Melbourne, Australia; abeer.mazher@csiro.au

Received: 18 December 2019; Accepted: 19 February 2020; Published: 21 February 2020

Abstract: Numerical modelling increasingly generates massive, high-dimensional spatio-temporal datasets. Exploring such datasets relies on effective visualization. This study presents a generic workflow to (i) project high-dimensional spatio-temporal data on a two-dimensional (2D) plane accurately (ii) compare dimensionality reduction techniques (DRTs) in terms of resolution and computational efficiency (iii) represent 2D projection spatially using a 2D perceptually uniform background color map. Machine learning (ML) based DRTs for data visualization i.e., principal component analysis (PCA), generative topographic mapping (GTM), t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP) are compared in terms of accuracy, resolution and computational efficiency to handle massive datasets. The accuracy of visualization is evaluated using a quality metric based on a co-ranking framework. The workflow is applied to an output of an Australian Water Resource Assessment (AWRA) model for Tasmania, Australia. The dataset consists of daily time series of nine components of the water balance at a 5 km grid cell resolution for the year 2017. The case study shows that PCA allows rapid visualization of global data structures, while t-SNE and UMAP allows more accurate representation of local trends. Furthermore, UMAP is computationally more efficient than t-SNE and least affected by the outliers compared to GTM.

Keywords: machine learning; spatio-temporal gridded datasets; dimensionality reduction; color maps; spatial visualization; quality assessment.

1. Introduction

One of the biggest challenges of the big data era is to make sense out of all the information available. Unfortunately, not all that huge volume of data is informative. Such datasets may contain spatial or temporal information or both spatial and temporal information. The information is available either in the form of grid or point data. However, gridded data is difficult to capture in low-dimensional space especially in Earth sciences, due to their dynamic and non-linear behavior.

Effective data visualization plays a key role in exploring such big datasets, finding patterns/features and outliers. Such insights are essential to develop hypotheses on the data-generating processes [1]. In addition, data visualization tools can help improve decision making, primary data analysis and information sharing [2]. Several visualization approaches exist in literature to extract the information based on graphs, charts, parallel coordinates and tree maps just to name a few [3,4]. So far, domain experts in Earth sciences rely on traditional visualization methods, such as maps and time series plots, to explore patterns and structures of high-dimensional spatio-temporal datasets. Data visualization is embedded in various inference and feature extraction techniques [5]. Therefore, visualizing

high-dimensional spatio-temporal datasets requires a dimensionality reduction step to extract the most informative feature dimensions [6].

Dimensionality reduction techniques (DRTs) in hydrology and hydrological modelling are mostly used as a precursor for classification and clustering multivariate datasets, such as hydrological time series [7,8]. Dimensionality reduction for spatial visualization of gridded data is, however, gaining increasing research interest [9,10] and have not been explored using emerging technologies.

Machine learning (ML) algorithms have allowed us to perform complex tasks with limited information in short amount of time and are able to represent large complex non-linear systems in a computationally efficient manner. In literature, there exist several ML based DRTs for data visualization [6,11–20]. The DRTs for data visualization are based on linear and non-linear techniques [6,11–14], spectral and stochastic embedding [15], parametric and non-parametric techniques [16–19], neighborhood preservation techniques [17], topographic mapping [18,19] and multi-dimensional scaling [20,21]. While these methods vary in their way, they preserve distances and neighborhood relationships between data points in the reduced dimensionality space, they all aim to minimize both the redundant information and loss of information.

Principal component analysis (PCA) is so far one of the oldest and the best-known DRT in data mining [13] and is widely used in hydrological sciences since 1960s [22]. PCA is a parametric linear technique that constructs a low-dimensional representation of a dataset by capturing maximum variation [16]. The alternative linear approaches include random projection [23,24] but they are not able to capture the non-linear structures in stochastic datasets. PCA was successfully used in past studies to identify variations in water quality parameters [25], understanding subsurface groundwater properties [26] and to explore different quality characteristics of water systems [27]. Although, PCA memory requirement is minimal i.e., only equal to the number of data points (P), however, the assumption of capturing only linear features limits the applicability of PCA.

There exist many non-linear extensions of PCA such as kernel PCA [28], manifold charting [29] and self-organizing maps (SOMs) [18]. SOM is an unsupervised neural network (NN) algorithm that performs a non-linear mapping of the dominant dependent features present in the high dimensional data to a low-dimensional grid [18,30].

On the other hand, generative topographic mapping (GTM) is a parametric non-linear technique first introduced as a probabilistic alternative to SOMs. GTM performs non-linear mapping from the latent space into the high dimensional data space and for data visualization, mapping is then inverted using Bayes' theorem, giving rise to a posterior distribution in latent space [19]. GTM overcomes many drawbacks of SOMs, such as it preserves the topological structure and retains the neighborhood information. The algorithm is used for various applications, such as in oil fraction determination from a mixture of oil, water and gas in a multi-phase pipeline [19], in mapping sparse data sequences to visualize the distribution of text-based documents [31], in classification of fault data [32] and in mapping the biopharmaceutical data [33]. Although GTM's computational memory requirement is assumed equivalent to P i.e., the number of data points, however, it comes with the cost of selecting the appropriate parameters, which may lead towards overfitting.

Multi-dimensional scaling (MDS) is the first non-parametric DRT that seeks a 2D representation of high-dimensional datasets, which preserve topology and distances [17]. Several extensions of MDS are available in literature, such as curvilinear component analysis [34] and curvilinear distance analysis [35], however, the capability to capture non-linear structures by MDS is limited and fine tuning of optimization parameters is required in the extended versions of MDS.

Many alternative non-parametric approaches are discussed in literature to capture non-linear structures, such as, Isotop [36] for preserving the neighborhood information but do not have any specific cost and objective functions. Another approach, the stochastic neighbor embedding (SNE) [37] is introduced with the explicit cost function along with the properties to preserve the neighboring information. Its main drawback is overcrowding of data points in the projected low dimensional space using gaussian distribution, which leads to the compact representation of dataset.

T-distributed SNE (t-SNE) [12] is an improved variation of SNE consists of a long-tailed distribution, hence large neighborhoods of the data can be matched by a wide range of scales in the two-dimensional (2D) projection and in this way avoid the overcrowding of data points. In fact, the t-SNE approach tries to match the probability distributions induced by the pairwise data dissimilarities in the original data space and the projected space. t-SNE has successfully been used to visualize high dimensional datasets to reveal local as well as global structures at several scales in various application areas such as, computer security [38], cancer biology [39], music analysis [40] and bioinformatics [41]. It is worth mentioning that t-SNE computational and memory complexity is quadratic in the number of data points (P^4) and the local nature of t-SNE makes it sensitive to the curse of inherent dimensionality of high-dimensional data [42]. In addition, t-SNE performance on the general dimensionality reduction tasks is still vague as it is primarily built for visualization purpose only.

Uniform manifold approximation and projection (UMAP) is a recently introduced non-parametric DRT, which shows its effectiveness in coping with diversity of dynamic and non-linear datasets [43]. It builds on strong mathematical foundations largely based on manifold theory and fuzzy topological representation, which allows it to scale to very large datasets in an efficient manner. Like t-SNE, UMAP has widely been used in the fields of bioinformatics [44], material science [45] and machine learning [46], however, so far, no application has been found in hydrology and the Earth sciences. UMAP computational efficiency equals P and has no computational restrictions on projected dimensions. This is because UMAP does not require global normalization. Further, UMAP is built on solid theoretical grounds useful for general purpose dimensionality reduction and preprocessing of machine learning techniques.

This study focuses on one linear and three non-linear ML based unsupervised DRTs for visualization i.e., principal component analysis (PCA) [16], generative topographic mapping (GTM) [19], t-distributed stochastic neighbor embedding (t-SNE) [12] and uniform manifold approximation and projection (UMAP) [43] summarized in Table 1 along with their respective computational efficiencies in terms of data points represented by P.

Table 1. Dimensionality reduction techniques and their respective computational efficiencies.

	Parametric	**Non-Parametric**
Linear	PCA (P)	-
Non-linear	GTM (P)	t-SNE (P^4), UMAP(P)

The reason to choose above mentioned DRTs is to test their practicality in terms of accuracy, resolution and computational efficiency for high dimensional spatio-temporal gridded dataset.

To quantify visualizations of selected DRTs, there exists various quality metrics in literature, either independent or dependent on DRTs. The quality is assessed by calculating pairwise proximities such as, distances, similarities/dissimilarities or probabilities between low-dimensional and high-dimensional space or by reproducing a high-dimensional space from a low-dimensional projection [47,48]. Different DRTs preserves different proximities, therefore, difficult to compare. The DRTs dependent quality metrics include neighborhood scales [49] and agreement evaluation criteria [50]. These evaluation criteria preserve distances between neighboring points and are not suitable for comparison of different pairwise proximities calculated from various DRTs. There are few DRT independent quality metrics, which includes scale independent [51], distance [52] and rank based criteria [53]. The scale independent criteria, such as Q_{NX} [51] use ranks instead of pairwise distances between the data points to define the nearest neighbors. The rank comparison-based approach i.e., co-ranking [53] has a benefit of comparing Euclidean distance of projected low-dimensional data points to any pairwise proximity of original high-dimensional data as ordering the neighboring points is possible in all cases. Although, the absolute information of a proximity is lost in the ranking procedure, however, the rank comparison-based technique is suitable for comparing different DRTs.

To visualize patterns effectively, the low-dimensional projected plane needs to be placed in a spatial context. In this regard, color maps will help the end users to explain spatio-temporal patterns intuitively. Visualizing higher dimensional data traditionally relies on directly mapping variables to R, G, B channels to create a pseudo-color image e.g., [54,55] or combining different variables into a predefined index such as, [56,57]. The main drawbacks of these methods are that they can only visualize limited number of variables and that they require an in-depth understanding of the data generating process to develop a meaningful index. High-dimensional data can be visualized by color coding data points according to their projection in a lower dimensional space [58]. One of the most challenging aspects of visualizing data through false color images is to develop a perceptually uniform color scheme in order not to inadvertently emphasize or obscure parts of the data range [59–61].

In crux, the process of dimensionality reduction incurs a loss of information. Where information loss is strongly linked with the preservation of geometry (distances, topology and reproducibility) and helps in evaluating the trustworthiness of visual maps as shown in [47] i.e., the greater the loss of quality, the less the preservation of geometry. Furthermore, the loss can be quantified to gain an in-depth insight into the visualization's accuracy. The critical analysis of the literature review reveals that computationally efficient and accurate methods are required, which automatically embeds dimensionality reduction in visualization workflow to develop hypothesis on complex non-linear systems.

The objective of this study is to suggest a workflow that, (i) projects the inherent structure of the high-dimensional spatio-temporal data in 2D accurately, (ii) Compare DRTs in terms of resolution and computational efficiency and (iii) spatially visualizes the structure in an intuitive manner through a perceptually uniform color coding of 2D reduced parameter space. Furthermore, to the best of our knowledge, such comparison hasn't been performed on hydrological dataset to extract local and global structures. These structures serve as a backbone to describe hypothetical phenomenon. There is a great need for such frameworks, which will help to reduce the uncertainty in observations along with the improved understanding of physics and dynamics of hydrological systems.

The next section describes the dataset used followed by the adopted methodology, which shows DRTs, accuracy metrics, color scheme adopted for visualization and computational efficiency comparison. Result section consists of different DRTs visualization maps along with their visualization quality quantification and time series analysis. Later, the discussion section will compare the DRTs in terms of accuracy, resolution and computational efficiency followed by the conclusion.

2. Materials and Methods

2.1. Dataset and Study Area

The Australian Water Resource Assessment (AWRA) model is an operational, near real-time continental landscape model [62]. For each day, the model partitions rainfall in millimeters (mm) per day into potential and actual evaporation, runoff and deep drainage to groundwater as well as the change in water storage in four soil compartments, expressed in mm.

These nine hydrological variables are calculated on a 5 × 5 km grid across Australia, considering variability in vegetation, topography and soil properties [63]. The ongoing model development is focused on improving the representation of local conditions and improving computational efficiency [64].

For this case study, we used the model outputs for the state of Tasmania from 1 January 2017 to 30 September 2017, downloaded from http://www.bom.gov.au/water/landscape on 30 September 2017 as shown in Figure 1. This dataset has 273 daily values for the nine hydrological variables on a spatial grid scale of ($-43.71 <$ latitude < -40.14) and ($144.49 <$ longitude < 149.41). The three-dimensional arrays of hydrological variable are normalized to the [0, 1] range individually and appended to each other to create a single three-dimensional array of size i.e., latitude by longitude by AWRA components × daily time series observations (72 by 79 by 9 × 273) as shown in Figure 2. This array is vectorized

and grid cells covering oceans are excluded. This results in a 2D array of size (72 × 79) by (9 × 273) for further processing i.e., dimensionality reduction and visualization.

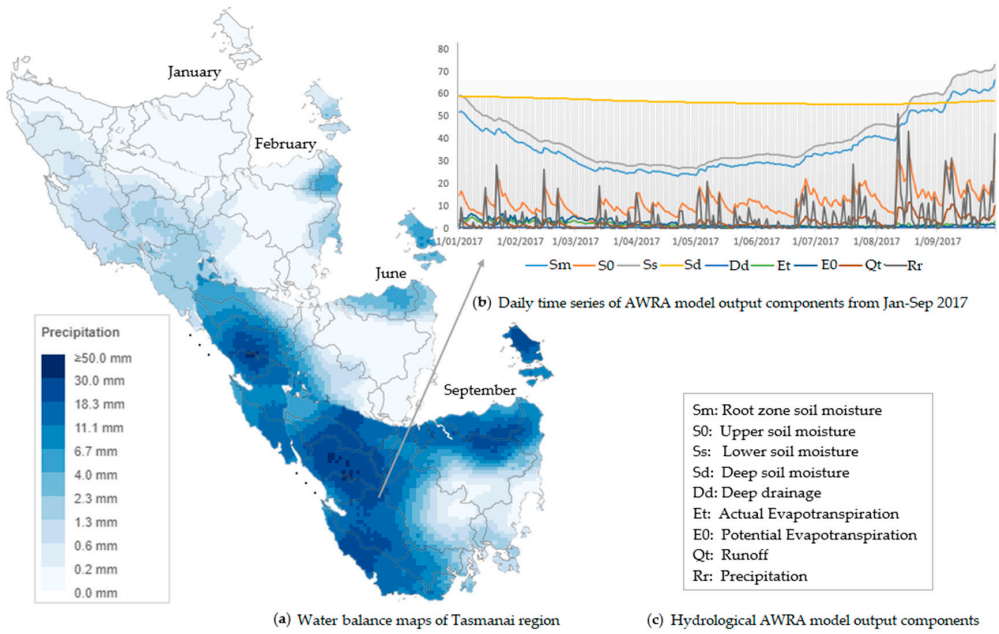

Figure 1. (**a**) Water balance map of Tasmania region for Australian Water Resource Assessment (AWRA) model output from January–September 2017, (**b**) each pixel shows daily time series of nine hydrological AWRA components from January–September 2017 and (**c**) hydrological components of AWRA model output.

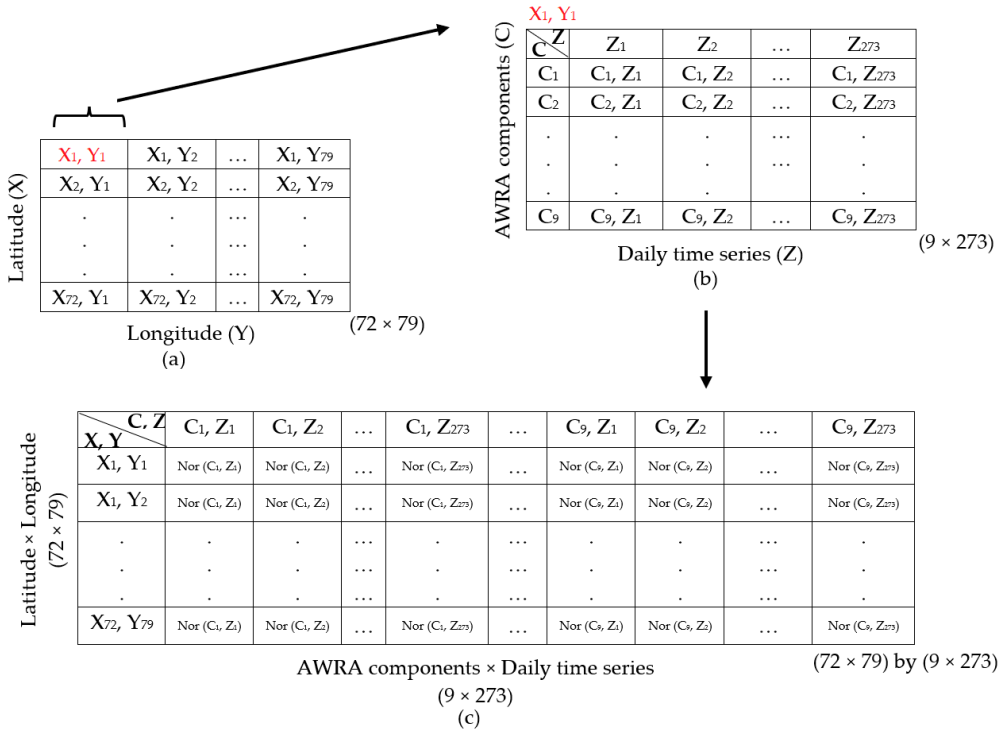

Figure 2. (**a**) Three-dimensional arrays of hydrological variables at the spatial grid scale of (−43.71 < latitude < −40.14) and (144.49 < longitude < 149.41) of size (72 × 79), (**b**) a spatial grid e.g., (X_1, Y_1) consists of (9 × 273) observations and (**c**) three-dimensional arrays are normalized (Nor) and append to each other to form a single three-dimensional array of size (72 × 79) by (9 × 273).

2.2. Methods

To visualize spatio-temporal structures locally and globally, four DRTs have been compared to efficiently and accurately visualize the multivariate hydrological AWRA model output components. Later, the perceptually uniform color scheme has been used, which allows user/domain experts to associate each data point on a 2D color plot by preserving its specific location on a spatial map of Tasmania. Figure 3 shows this workflow.

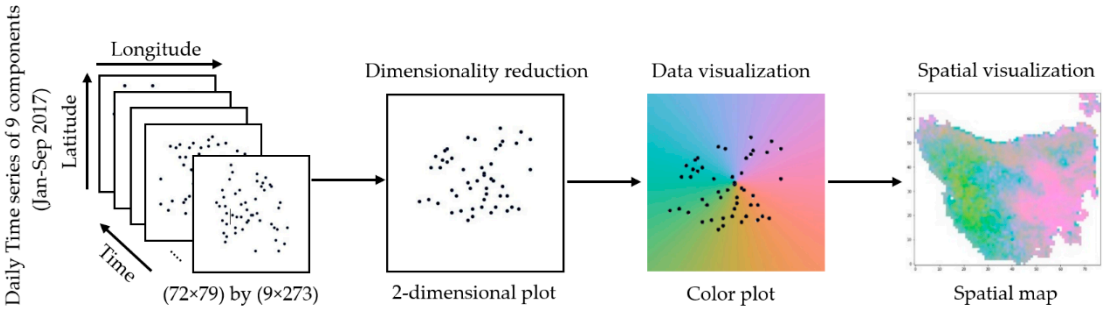

Figure 3. Visualization workflow for high-dimensional spatio-temporal gridded datasets.

2.2.1. Dimensionality Reduction Techniques (DRTs)

The objective of dimensionality reduction techniques is to find a low-dimensional projection of a high-dimensional dataset with minimal loss of the information [6]. These techniques visualize correlations and patterns in high-dimensional spatio-temporal datasets.

For this case study, two parametric (PCA, GTM) and two non-parametric (t-SNE, UMAP) DRTs are compared.

PCA is a parametric linear technique that projects a dataset onto a 2D space defined by linearly uncorrelated principal components [16].

Another parametric but non-linear DRT used for visualization is GTM [19,65]. GTM aims to extract a low-dimensional representation of dataset, initially unknown hence called latent, lies on high-dimensional data space. Such mapping uses non-linear function to map points in latent space corresponding to points in data space, provides a map with the distribution of data points centered at lattice. A lattice consists of grid nodes window. These lattices have an attached responsibility i.e., either mean or mode of a distribution, and are used to provide visualization of the map for individual data points in 2D latent space. The size of the lattice is dependent on the chosen width factor of radial basis function, which has a direct influence on the visualization. Smaller the lattice size, less compact visualization can be attained. The hyperparameters govern the execution of GTM including (i) number of nodes used for tuning the GTM resolution (k); (ii) number of hidden units in radial basis function (m); (iii) width factor of radial basis function (s) and (iv) regularization coefficient (r) [66].

Unlike PCA and GTM, t-SNE is a non-parametric non-linear technique [12]. t-SNE visualizes high-dimensional data by giving each datapoint a location in a low-dimensional projected map. It calculates the probability of similarity between data points using gaussian distribution in high-dimensional space and calculates the same for its corresponding data points in low-dimensional space using T-distribution. The similarity of data points is calculated as conditional probabilities i.e., the points nearest to a defined center are picked by gaussian distribution and T-distribution in high and low dimensional space, respectively. Later, t-SNE tries to minimize the difference between the conditional probabilities in high-dimensional and low-dimensional space for the best possible visualization of data points in 2D using gradient descent method. The objective is to preserve the neighborhood information without any pre-requisite of user defined input.

UMAP is also a non-parametric non-linear DRT and it searches for a low-dimensional projection of a dataset that has the closest possible equivalent fuzzy topological structure (made up of local

manifold approximations which preserve distances) in high-dimensional space [44]. UMAP then minimizes the cross-entropy between low and high-dimensional space to optimize the visualization in low-dimensional space. UMAP preserves the essential topological structure of the learned manifold in the 2D representation of a dataset. Three main input parameters need to be defined by the user i.e., (i) the number of neighboring points used in local approximation of manifold structure (n), which ranges from 5 to 50; (ii) factor (d) with values between 0.001 to 0.5 controlling how tightly the embedding is allowed to compress points together and (iii) the choice of metric (c) used to measure distance in the input space including minkowski style metrics, spatial metrics, angular and correlation metrics [67].

The DRTs are performed in Python 3.6 using sklearn package for PCA and t-SNE. GTM is applied using the ugtm package imported from https://github.com/hagax8/ugtm and UMAP is applied using the umap package imported from https://github.com/lmcinnes/umap.

2.2.2. Quality Metric

The Q_{NX} measure [51] is used to quantify DRT based visualizations in a topology preserving manner i.e., how well neighborhood relationship can be preserved between data vectors to capture in 2D. Q_{NX} relies on the ranks of sorted distances between the high-dimensional and 2D projections. Q_{NX} measure is independent of any DRT, therefore, successfully used for the quantification of 2D projected visualizations for comparative analysis. This technique averages the quality curve Q_{NX} over varying values of K-ary of neighborhood (K) [51] as shown in Equation (1).

$$Q_{NX}(K) = \frac{1}{KN} \sum_{k=1}^{K} \sum_{l=1}^{K} Q_{kl} \quad (1)$$

$$Q_{kl} = \left\{ (i,j) \rho_{ij} = k \text{ and } r_{ij} = l \right\} \quad (2)$$

In Equation (1) Q_{NX} is a quality metric ranging between [0, 1], 1 means a perfect projection and vice versa. N is the total number of observations in high-dimensional space i.e., (72×79) by (9×273), X represents projected low-dimensional vectors and K is the neighborhood size. In Equation (2), ρ_{ij} is the rank of a sorted distance (ξ_i, ξ_j) in a high-dimensional space corresponding to a rank r_{ij} allotted to a sorted distance (x_i, x_j) in 2D projection. A more comprehensive overview of the mathematical details is provided in [46,51].

The above quantification criteria show a major advancement over the distance preservation measurement as the use of ranks allow distances to grow or shrink, makes it scale-independent, given that their orders do not change. Such criteria only dependent on neighborhood size K, produces curve that may be analyzed on various scales.

2.2.3. Visualization

The next step in the workflow is to super impose the perceptually uniform color scheme on the projected 2D plane to generate a spatial map of the dataset in which points with similar colors indicate similar data vectors [57]. The color scheme is based on HSL_{UV} (www.hsluv.org), a human-friendly alternative to the Hue, Saturation, and Lightness (HSL) color space. It extends the perceptually uniform CIELUV color space with a saturation component that allows chroma to be expressed as a percentage. It is to be anticipated that the combination of three components i.e., colored scheme, colored map and colored time series in a machine learning based visualization workflow allows to capture rich structure of the data and display results in a format domain expert are familiar with.

3. Results

To accurately present the spatio-temporal structure of high-dimensional AWRA model output in 2D and to visualize it using the perceptually uniform color scheme, the comparative analysis has been performed using PCA, GTM, t-SNE and UMAP.

The color plots corresponding to their spatial maps are provided in Figure 4a–c respectively. The background color to the plot allows us to associate each AWRA pixel on a spatial map with a position on the color plot. Each pixel consists of nine time series and reflect the water balance in a specific location.

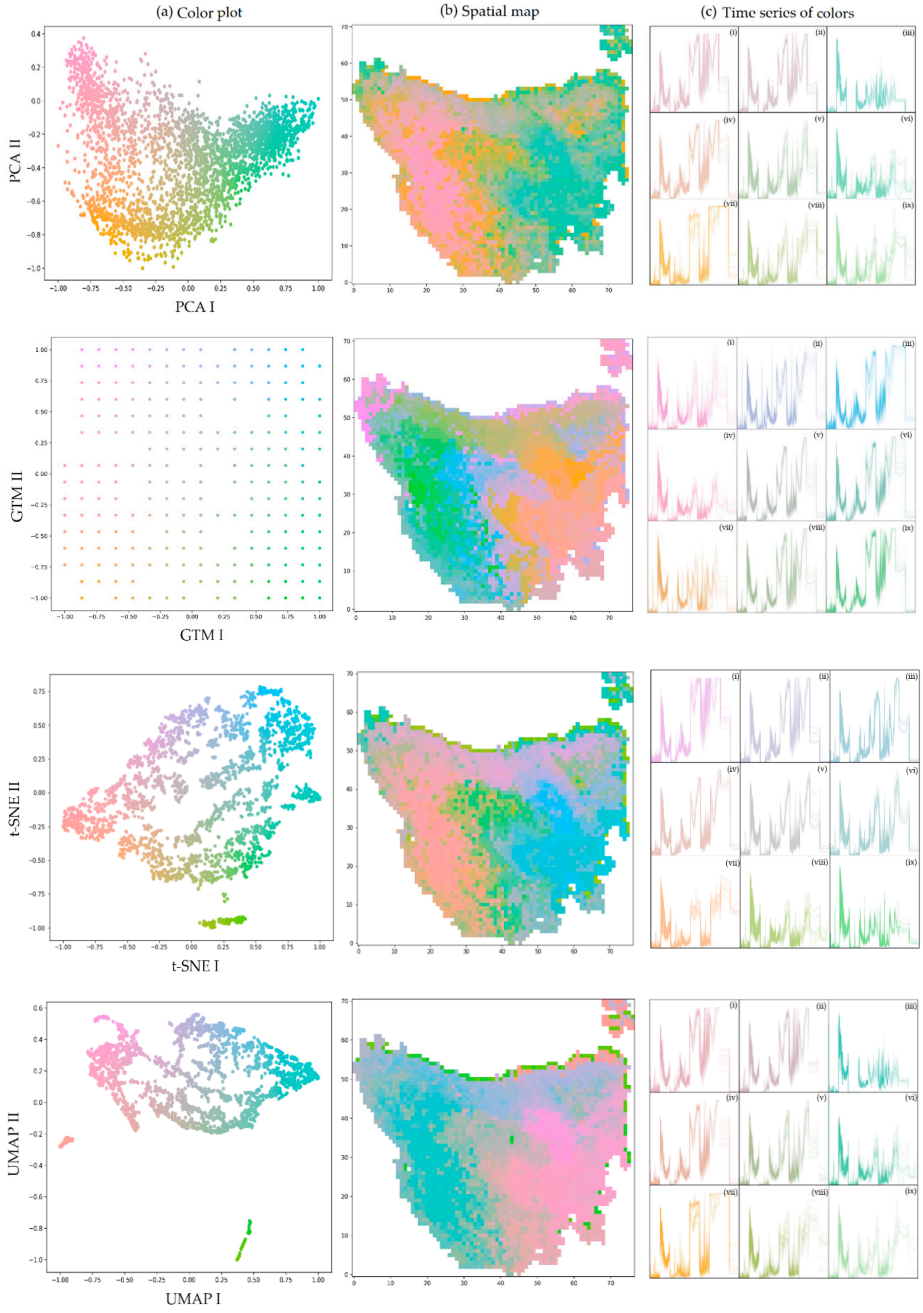

Figure 4. For principal component analysis (PCA), generative topographic mapping (GTM), t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP), (**a**) color plot shows a projection of high-dimensional data on 2D plane by preserving topology and maximize distances between projected and high-dimensional data for AWRA model output for Tasmania, 2017, (**b**) spatial map of Tasmania provides a quick spatial visualization of high-dimensional data features and (**c**) time series colored according to their position in 2D space. The subplots show 20 randomly selected time series from within 9 equally sized regions of the 2D space.

In Figure 4a PCA, t-SNE and UMAP are one-to-one mapping, where each data point is represented in a low-dimensional 2D space. GTM however is a many-to-one mapping in which multiple data points can be mapped to a single node, which represents the mean of a probability distribution. The best GTM visualization for the dataset in hand consists of a parameter selection (k = 16, m = 10, s = 0.3, r = 0.1) along with the mean data points representation. Further the choice of parameters (n = 10, d = 0.1 and c = correlation metric) provides the best visualization results for UMAP.

Figure 4b visualizes the main trends in the dataset arising from topographic and rainfall gradients. As the color scheme is designed to gradually blend from one color to another, colors close to each other indicate similar behavior. This allows us to quickly visualize areas with similar behavior forming clusters and find outlying values. In this regard, t-SNE and UMAP, appears to provide higher contrast to identify subtle trends and dissimilarities. This is largely due to the ability of t-SNE to capture small pair-wise distances, i.e., local structures in a 2D projected map, which resulted in better and clear local patterns.

Further UMAP performance is visually comparable to t-SNE but arguably capture mores of the global structure than local structure. This is due to the tradeoff between the number of neighborhoods (n) chosen in local approximations of manifold structure and the reasonable tightness of data points embedding (d) in optimizing the visualization quality. As larger n with tight d will average out the local approximations in the process and result in potentially densely packed regions, which captures the global structure better. GTM performs relatively poor in capturing any type of trends largely due to its dependence on the efficient parameter selection and sensitivity to outliers. Furthermore, the reason may be its dependence on the mean nodes of the data points instead of the data points itself. The first two components of PCA summarize the main trends in the data, which in results in displaying more global patterns than other compared techniques.

It is noticeable that without applying any clustering technique, the continuous perceptual uniform color plot scheme provides an in-depth insight into different groupings with an intuitive interpretation and meaningful patterns.

Figure 4(c(i)–(ix)) can further authenticate the above given statements. The 2D space is divided in 9 equal regions and within each region, 20 points are randomly selected. The associated time series are plotted with the corresponding color. The time series of colors associated with t-SNE and UMAP are more similar within a 2D space region, showing less variation in changing colors abruptly and results in forming clusters more accurately at local scale; whereas, the time series associated with PCA and GTM show more variation in time series within a 2D space region. These variations are clearly shown by varying time series colors in Figure 4(c(i)–(ix)), resulted in identifying patterns locally with less accuracy.

Figure 5 shows the Q_{NX} metric for DRTs. The relatively constant Q_{NX} values across neighborhood values is an indication that t-SNE and UMAP capture the local structures of high-dimensional spatio-temporal dataset in 2D projection as well as global structures. However, UMAP performance to t-SNE is slightly better at capturing the global trends but the local trends are affected by the outliers/noise. Furthermore, GTM performs poor in capturing the local as well as global trends due to its sensitivity to the outliers. The first two principal components, not surprisingly, do not provide much insight into the local structures, but perform better than t-SNE and UMAP to capture global structures as shown by higher values of Q_{NX} for larger neighborhoods (K) shown in Figure 5.

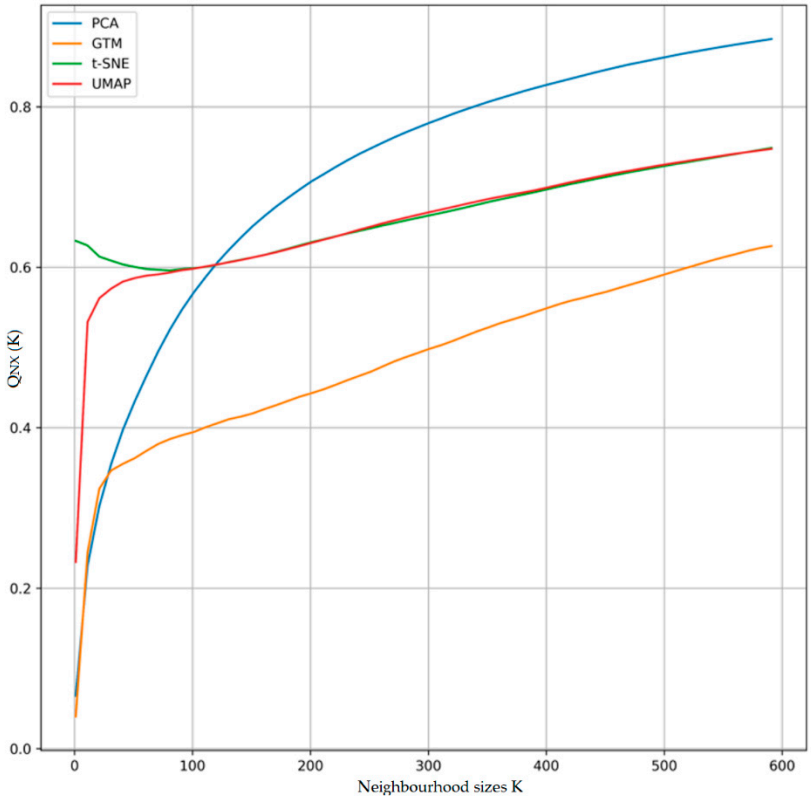

Figure 5. Q_{NX} visualization quantification curves PCA, GTM, t-SNE and UMAP. PCA provides better global patterns (K > 120), whereas, GTM performance is poor in comparison. Further, t-SNE captures the local structures better (K < 120) and UMAP captures better global structure between (250 < K < 550).

Table 2 further justifies Figure 5 by providing the Q_{NX} metric values against their varying neighborhood sizes K for four DRTs. As stated before, if $Q_{NX} \sim 1$, it shows perfect projection and vice versa. Drastic increase in Q_{NX} values against larger K for PCA compared to GTM, t-SNE and UMAP shows that PCA is better at capturing global trends as Q_{NX} values quickly approaches to 1. However, Q_{NX} values against varying values of K are far from 1 for GTM, therefore, does not capture any kind of trend with higher accuracy.

Table 2. Visualization quantification index (Q_{NX}) for DRTs against neighborhood sizes (K).

Q_{NX}(K)/K	0	60	120	180	240	300	360	420	480	540
PCA	0.06	0.50	0.60	0.70	0.75	0.79	0.81	0.83	0.85	0.86
GTM	0.04	0.38	0.42	0.44	0.47	0.50	0.52	0.57	0.59	0.61
t-SNE	0.63	0.60	0.60	0.62	0.65	0.67	0.68	0.70	0.72	0.73
UMAP	0.22	0.59	0.60	0.62	0.652	0.673	0.682	0.702	0.722	0.732

On the other hand, t-SNE showed better performance in capturing local as well as global trends. This is due to higher Q_{NX} values against K compared to GTM and UMAP as shown in Table 2. It is important to note that, UMAP is slightly better at capturing global trends compared to t-SNE for the range of K, i.e., 250 < K < 550.

As far as the computational efficiency is concerned, PCA took 20 seconds to run, whereas GTM takes at least 150 seconds to run. On the other hand, UMAP is much faster in producing results due to its non-parametric nature and took approximately 15 seconds, however, a non-parametric t-SNE took only 80 seconds to run. The experiment is performed on 64-bit operating system with 32 GB Ram.

4. Discussion

Choosing an appropriate data representation method is not a trivial task as they differ in the goals of exploring either correlations, clusters or patterns in datasets with varying computational efficiencies. Most of the DRTs are designed for high-dimensional point datasets. The gridded spatio-temporal datasets are computationally more challenging.

Generally speaking, linear techniques compared to non-linear DRTs are more flexible in representing high-dimensional structures in lower dimensions and therefore will incur less loss of information. Topographic mappings and spectral embedding are designed to preserve the topography or geometry of the dataset, whereas neighborhood preservation techniques, such as multi-dimensional scaling and neural networks, aim to retain the multi-dimensional distances between data points in the low-dimensional projections. Furthermore, the parametric DRTs optimize the parameters in the process of training, which will provide out-of-sample extensions, whereas the non-parametric DRTs do not.

Specific to the linear parametric DRT discussed in this case study i.e., PCA, a linear projection cannot faithfully reveal the non-linear structures of the datasets and disturbs the local neighborhood structures. Further, for very high-dimensional datasets, PCA becomes costly and sensitive to noise due to its dependency on data covariance matrix. However, PCA is a useful preprocessing step for very high-dimensional datasets to later proceed with the non-linear feature extraction techniques. Alternatively, non-linear features are well captured by the parametric DRT i.e., GTM, however, requires an appropriate selection of parameters.

On the other hand, non-parametric techniques have an advantage of fast processing and do not assume any functional form of mapping to regulate parameters, however, suffers from a disadvantage of not providing out of sample extension. T-SNE and UMAP are the main non-parametric ML based DRTS discussed in this case study. UMAP is preferable to use for general purpose DRT, however, t-SNE is preferred for visualization.

This case study suggested a workflow by comparing four ML based DRTs in terms of accuracy, efficiency and resolution suitable for high-dimensional spatio-temporal gridded datasets followed by its visualization quantification.

PCA retains large pairwise distances in the reduced dimensional space defined by the first two principal components only. Local structures may be captured in other principal components. t-SNE performs well for the high dimensional gridded datasets as its non-linear mapping function helps to capture various spatio-temporal structures efficiently. It is important to mention here that UMAP is competitive to t-SNE for visualization quality, however, preserve more of a global structure compared to t-SNE. Furthermore, the performance of GTM largely depends on its parameter selection i.e., number of nodes (k). With the increase in m, the lower-dimensional space points get more clustered together due to the network overtraining, resulting in a more precise visualization rather than representing optimized structures.

Overall, UMAP is a general-purpose DRT and can be recommended to treat high-dimensional datasets with superior run time performance to capture non-linear global structure more accurately compared to t-SNE. UMAP and t-SNE can both handle large non-linear datasets more efficiently, however, non-linear GTM due to its parametric nature is prone to over-fitting.

All above discussed DRTs for data visualization have some associated advantages and disadvantages, however, will assist domain experts to select suitable technique for their dataset. Identifying the inherent structure of spatio-temporal dataset will, however, always be hampered by the information loss that is unavoidable when representing high-dimensional data in two dimensions. The selection of suitable ML based DRT depends on the dataset in hand. If the nature of dataset is linear than PCA is best at capturing the local and global features, however, non-linear datasets can be handled well by parametric GTM. If the parameters are more difficult to decide than UMAP and t-SNE are preferable to capture non-linear trends. However, it should be kept in mind that t-SNE is computationally more expensive compared to UMAP.

Moreover, different DRTs often result in very different visualizations and it is hard to decide the best suitable DRT for a given dataset at hand. It is often not clear if differences in the visualizations are due to the data structure or the method chosen. Several techniques have been developed to assess the quality of low dimensional projected visualizations. However, the Q_{NX} quality metric is suggested to quantify visualizations as it does not depend on any DRT. Further, to visualize spatial patterns of a single quantity of interest the perceptually uniform color scheme is recommended in order to capture patterns in a diverse range.

5. Conclusions

The suggested workflow applied DRTs to visualize the multivariate AWRA model output hydrological components in order to determine the prominent spatio-temporal features followed by its quantification to access the visualization accuracy. The comparative analysis of four DRTs i.e., PCA, GTM, t-SNE and UMAP by accounting a perceptual uniform color scheme has been performed on AWRA model output for the Tasmania region.

t-SNE and UMAP are effective in detecting local spatial patterns with high resolution as compared to the GTM, whereas, GTM lacks resolution in explaining the local as well as global trends. On the other hand, PCA is better at detecting the global patterns. The time-series color plot further validates the results and the quality metric Q_{NX} proves it quantitatively. Moreover, t-SNE proves to be computationally quite expensive for high-dimensional spatio-temporal datasets, compared to PCA, UMAP and GTM but provides much better insight almost equivalent to UMAP in explaining the data structures and patterns. Furthermore, GTM is much sensitive to outliers compared to UMAP.

In essence, t-SNE and UMAP are better to use when non-linear trends are expected and local trends are of much more importance, however, the UMAP is computationally more efficient. Furthermore, PCA can capture global trends better for linear datasets, whereas, the parametric nature of the GTM to capture non-linear trends makes it harder to capture any kind of trend.

The suggested workflow is beneficial for the exploratory data analysis of hydrological data.

Funding: The APC was funded by Deep Earth Imaging-Future Science Platform, CSIRO, Australia.

Acknowledgments: I would like to thank Luk Peeters (CSIRO, Deep Earth Imaging-Future Science Platform) contribution in concept designing and refining the manuscript.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Keim, D.; Kohlhammer, J.; Ellis, G.; Mansmann, F. Visual Analytics. In *Mastering the Information Age: Solving Problems with Visual Analytics*; Eurographics Association: Munich, Germany, 2010; pp. 7–18.
2. Sucharitha, V.; Subash, S.R.; Prakash, P. Visualization of Big Data: Its Tools and Challenges. *Int. J. Appl. Eng. Res.* **2014**, *9*, 5277–5290.
3. Kerren, A.; Stasko, J.T.; Fekete, J.D.; North, C. *Information Visualization—Human-Centered Issues and Perspectives*; Springer: Berlin/Heidelberg, Germany, 2008.
4. Güler, C.; Thyne, G.D.; McCray, J.E. Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. *Hydrogeol. J.* **2002**, *10*, 455–474. [CrossRef]
5. Ward, M.; Grinstein, G.; Keim, D.A. *Interactive Data Visualization: Foundations, Techniques, and Application*; A K Peters, Ltd.: Boca Raton, FL, USA, 2010.
6. Van der Maaten, L.J.P.; Postma, E.O.; van den Herik, H.J. *Dimensionality Reduction: A Comparative Review*; Tilburg University: Tilburg, The Netherland, 2009.
7. Kennard, M.J.; Pusey, B.J.; Olden, J.D.; MacKay, S.J.; Stein, J.L.; Marsh, N. Classification of natural flow regimes in Australia to support environmental flow management. *Freshw. Biol.* **2010**, *55*, 171–193. [CrossRef]
8. Herbst, M.; Gupta, H.V.; Casper, M.C. Mapping model behaviour using {S}elf-{O}rganizing {M}aps. *Hydrol. Earth Syst. Sci.* **2009**, *13*, 395–409. [CrossRef]

9. Wang, N.; Biggs, T.W.; Skupin, A. Com puters, Environ ment and Urban System s Visualizing gridded time series data with self-organizing maps: An application to multi-year snow dynamics in the Northern Hemisphere. *Comput. Environ. Urban Syst.* **2013**, *39*, 107–120. [CrossRef]
10. Biswas, A.; Lin, G.; Liu, X.; Shen, H.W. Visualization of Time-Varying Weather Ensembles across Multiple Resolutions. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 841–850. [CrossRef] [PubMed]
11. Gisbrecht, A.; Hammer, B. Data visualization by nonlinear dimensionality reduction. *Wires Data Min. Knowl. Discov.* **2015**, *5*, 51–73. [CrossRef]
12. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
13. Lee, J.A.; Verleysen, M. *Nonlinear Dimensionality Reduction*; Springer: New York, NY, USA, 2007.
14. Bunte, K.; Biehl, M.; Hammer, B. A general framework for dimensionality-reducing data visualization mapping. *Neural Comput.* **2012**, *24*, 771–804. [CrossRef]
15. Saul, L.K.; Roweis, S.T. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *J. Mach. Learn. Res.* **2003**, *4*, 119–155.
16. Jolliffe, I.T. *Principal Component Analysis*, 2nd ed.; Springer: New York, NY, USA, 2002.
17. Nam, K.; Je, H.; Choi, S. Fast Stochastic Neighbor Embedding: A trust-region algorithm. In Proceedings of the IEEE International Joint Conference on Neural Networks, Budapest, Hungary, 25–29 July 2004.
18. Kohonen, T. *Self-Organizing Maps*; Springer: Heidelberg, Germany, 2001.
19. Bishop, C.M.; Svensen, M. GTM: The Generative Topographic mapping. *Neural Comput.* **1998**, *10*, 215–234. [CrossRef]
20. Cox, T.F.; Cox, M.A.A. *Multidimensional Scaling*; Chapman & Hall: London, UK, 1994.
21. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef] [PubMed]
22. James, R.W. Multivariate Statistical Methods in Hydrology-A comparison using data of known functional relationship. *Water Resour. Res.* **1965**, *1*, 447–461.
23. Kaski, S. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In Proceedings of the International Joint Conference on Neural Networks (IJCNN'98), Anchorage, AK, USA, 4–9 May 1998.
24. Bingham, E.; Mannila, H. Random projection in dimensionality reduction: Applications to image and text data. In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001), San Francisco, CA, USA, 26–29 August 2001.
25. Atlas, L.; Isik, M.; Kavurmaci, M. Determination of arsenic levels in the water resources of Aksaray Province, Turkey. *J. Environ. Manag.* **2011**, *92*, 2182–2192.
26. Mahlknecht, J.; Steinich, B.; Navarro de León, I. Groundwater chemistry and mass transfers in the independence aquifer, central Mexico, by using multivariate statistics and mass-balance models. *Environ. Geol.* **2004**, *45*, 781–795. [CrossRef]
27. Azhar, S.C.; Aris, A.Z.; Yusoff, M.K.; Ramli, M.F.; Juahir, H. Classification of River Water Quality Using Multivariate Analysis. *Procedia Environ. Sci.* **2015**, *30*, 79–84. [CrossRef]
28. Schölkopf, B.; Smola, A.J.; Müller, K.R. Kernel principal component analysis. In *Advances in Kernel Methods*; Support Vector Learning: Lausanne, Switzerland, 1999.
29. Brand, M. Charting a manifold. In *Advances in Neural Information Processing Systems 15*; MIT Press: Cambridge, MA, USA, 2003.
30. Yin, H. On multidimensional scaling and the embedding of self-organising maps. *Neural Netw.* **2008**, *21*, 160–169. [CrossRef]
31. Kaban, A. A scalable generative topographic mapping for sparse data sequences. In Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05), Las Vegas, NV, USA, 4–6 April 2005.
32. Zhong, F.; Zheng, X.; Tan, Z.; Shi, T. Application of generative topographic mapping to the classification of bearing fault. In Proceedings of the IEEE International Conference on Control and Automation, Guangzhou, China, 30 May–1 June 2007.
33. Gaspar, H.A.; Marcou, G.; Horvath, D.; Arault, A.; Lozano, S.; Vayer, P.; Varnek, A. Generative topographic mapping-based classification models and their applicability do-main: Application to the biopharmaceutics drug disposition classification system (BD-DCS). *J. Chem. Inf. Model.* **2013**, *53*, 3318–3325. [CrossRef]

34. Demartines, P.; Hérault, J. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Trans. Neural. Netw.* **1997**, *8*, 148–154. [CrossRef]
35. Lee, J.A.; Lendasse, A.; Verleysen, M. Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis. *Neurocomputing* **2004**, *57*, 49–76. [CrossRef]
36. Lee, J.A.; Verleysen, M. Nonlinear projection with the isotop method. Proceedings of International Conference on Artificial Neural Networks (ICANN'2002), Madrid, Spain, 28–30 August 2002.
37. Hinton, G.; Roweis, S. Stochastic neighbor embedding. In *Processing of Advances in Neural Information Systems (NIPS)*; MIT Press: Vancouver, CO, Canada, 2002.
38. Gashi, I.; Stankovic, V.; Leita, C.; Thonnard, O. An experimental study of diversity with off the-shelf antivirus engines. In Proceedings of the Eighth IEEE International Symposium on Networking Computing and Applications (NCA 2009), Cambridge, MA, USA, 9–11 July 2009.
39. Abdelmoula, W.M.; Balluff, B.; Englert, S.; Dijkstra, J.; Reinders, M.J.T.; Walch, A.; McDonnell, L.A.; Lelieveldt, B.P.F. Data-driven identification of prognostic tumor subpopulations using spatially mapped t-sne of mass spectrometry imaging data. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 12244–12249. [CrossRef] [PubMed]
40. Hamel, P.; Eck, D. Learning features from music audio with deep belief networks. In Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), Utrecht, The Netherlands, 9–13 August 2010.
41. Wallach, I.; Lilien, R. The protein-small-molecule database (psmdb), a non-redundant structural resource for the analysis of protein-ligand binding. *Bioinformatics* **2009**, *25*, 615–620. [CrossRef] [PubMed]
42. Bengio, Y. *Learning Deep Architectures for AI*; Universite de Montreal: Montreal, QC, Canada, 2007.
43. Cornell University. Statistics, Machine Learning. Available online: https://arxiv.org/abs/1802.03426 (accessed on 10 March 2018).
44. Cold Spring Harbor Laboratory. The Preprint Service for Biology. Available online: https://www.biorxiv.org/content/10.1101/298430v1 (accessed on 15 May 2018).
45. Fuhrimann, L.; Moosavi, V.; Ohlbrock, P.O.; Dacunto, P. Data-driven design: Exploring new structural forms using machine learning and graphic statics. In Proceedings of the IASS Annual Symposium (IASS 2018), Boston, USA, 16–20 July 2018.
46. Cornell University. Computer Science, Machine Learning. Available online: https://arxiv.org/abs/1810.03052 (accessed on 10 October 2018).
47. Gracia, A.; Gonzalez, S.; Robles, V.; Menasalvas, E. A methodology to compare dimensionality reduction algorithms in terms of loss of quality. *Inf. Sci.* **2014**, *270*, 1–27. [CrossRef]
48. Lee, J.A.; Verleysen, M. Quality assessment of nonlinear dimensionality reduction based on K-ary neighborhoods. In Proceedings of the workshop and conference on New Challenges for Feature Selection in Data Mining and Knowledge Discovery, Antwerp, Belgium, 15 September 2008.
49. Mokbel, B.; Lueks, W.; Gisbrecht, A.; Hammer, B. Visualizing the quality of dimensionality reduction. *Neurocomputing* **2013**, *112*, 109–123. [CrossRef]
50. Gorban, A.; Zinovyev, A. Principal manifolds and graphs in practice: From molecular biology to dynamical systems. *Int. J. Neural. Syst.* **2010**, *20*, 219–232. [CrossRef]
51. Lee, J.; Verleysen, M. Scale-independent quality criteria for dimensionality reduction. *Pattern Recogn. Lett.* **2010**, *31*, 2248–2257. [CrossRef]
52. Venna, J.; Peltonen, J.; Nybo, K.; Aidos, H.; Kaski, S. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *J. Mach. Learn. Res.* **2010**, *11*, 451–490.
53. Lee, J.; Verleysen, J. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing* **2009**, *72*, 1431–1443. [CrossRef]
54. Kennett, B.; Chopping, R.; Blewett, R. *The Australian Continent: A Geophysical Synthesis*; ANU Press and Geoscience Australia: Acton, Australia, 2018.
55. Minty, B.; Franklin, R.; Milligan, P.; Richardson, M.; Wilford, J. The Radiometric Map of Australia. *Explor. Geophys.* **2009**, *40*, 325. [CrossRef]
56. Gao, B. NDWI-A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sens. Environ.* **1996**, *58*, 257–266. [CrossRef]
57. Peeters, L. A Background Colour Scheme for Piper Plots to Spatially Visualize Hydro-chemical Patterns. *Groundwater* **2013**, *52*, 2–6. [CrossRef] [PubMed]

58. Peeters, L.; Bação, F.; Lobo, V.; Dassargues, A. Exploratory data analysis and clustering of multivariate spatial hydrogeological data by means of GEO3DSOM, a variant of Kohonen's Self-Organizing Map. *Hydrol Earth Syst. Sci.* **2007**, *11*, 1309–1321. [CrossRef]
59. Bujack, R.; Turton, T.L.; Samsel, F.; Ware, C.; Rogers, D.H.; Ahrens, J. The Good, the Bad, and the Ugly: A Theoretical Framework for the Assessment of Continuous Colourmaps. *IEEE Trans. Vis. Comput. Graph.* **2018**, *24*, 923–933. [CrossRef] [PubMed]
60. Cornell University. Computer Science, Graphics. Available online: https://arxiv.org/abs/1509.03700 (accessed on 18 June 2018).
61. Light, A.; Bartlein, P.J. The end of the rainbow? Colour schemes for improved data graphics. *Eos Trans. AGU* **2004**, *85*, 385–391. [CrossRef]
62. Vaze, J.; Viney, N.; Stenson, M.; Renzullo, L.; Van Dijk, A.; Dutta, D.; Crosbie, R.; Lerat, J.; Penton, D.; Vleeshouwer, J.; et al. The Australian Water Resource Assessment System (AWRA). In Proceedings of the 20th International Congress on Modelling and Simulation (MODSIM2013), Adelaide, Australia, 1–6 December 2013.
63. Peeters, L.J.M.; Crosbie, R.S.; Doble, R.C.; Van Dijk, A.I.J.M. Conceptual evaluation of continental land-surface model behaviour. *Environ. Modell. Softw.* **2013**, *43*, 49–59. [CrossRef]
64. Gladish, D.W.; Pagendam, D.E.; Peeters, L.J.M.; Kuhnert, P.M.; Vaze, J. Emulation Engines: Choice and Quantification of Uncertainty for Complex Hydrological Models. *J. Agric. Biol. Envirostats.* **2017**, *23*, 9–62. [CrossRef]
65. Kireeva, N.; Baskin, I.I.; Gaspar, H.A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modelling and Dataset Comparison. *Mol. Inf.* **2012**, *31*, 301–312. [CrossRef]
66. Gaspar, H.A. ugtm: A Python Package for Data Modeling and Visualization Using Generative Topographic Mapping. *J. Open Res. Softw.* **2018**, *6*, 26. [CrossRef]
67. Basic UMAP Parameters. Available online: https://umap-learn.readthedocs.io/en/latest/parameters.html (accessed on 20 October 2018).