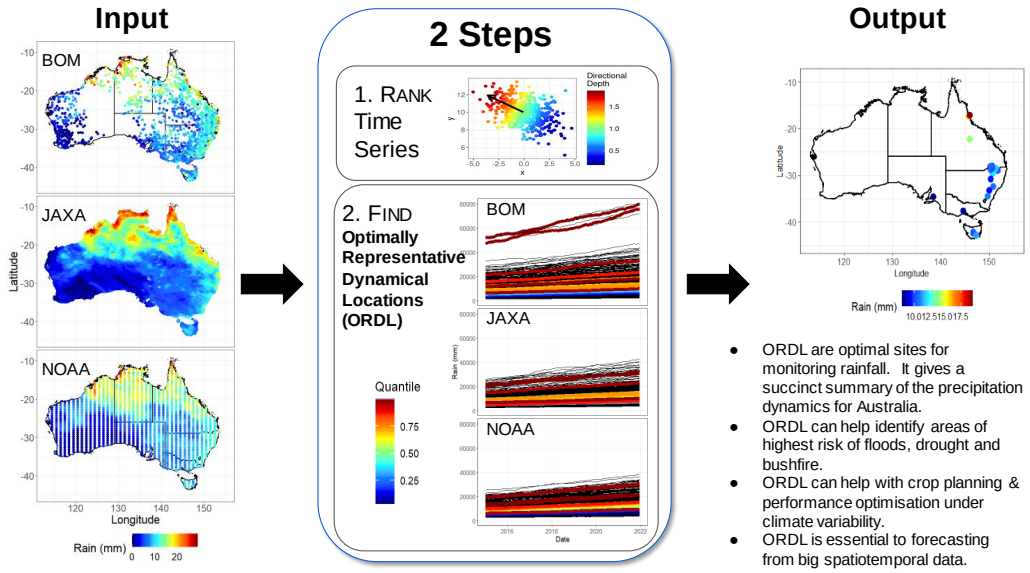
Graphical Abstract

Discovering optimally representative dynamical locations (ORDL) in big multivariate spatiotemporal data: a case study of precipitation in Australia from space to ground sensors

Benjamin Hines, Guoqi Qian*, Antoinette Tordesillas



Highlights

Discovering optimally representative dynamical locations (ORDL) in big multivariate spatiotemporal data: a case study of precipitation in Australia from space to ground sensors

Benjamin Hines, Guoqi Qian*, Antoinette Tordesillas

- Adaptation of statistical depth to give a more quantile like description of multivariate data by using eigenvalue decomposition.
- Generalisation of the univariate empirical dynamical quantiles to a multivariate setting using the adaptation of the statistical depth function.
- Multivariate empirical dynamical quantiles are used to find ORDL in a big multi-source spatiotemporal data on precipitation.
- ORDL highlight extreme variability in Queensland and new rain gauge locations to capture precipitation dynamics in Australia under climate change.

Discovering optimally representative dynamical locations (ORDL) in big multivariate spatiotemporal data: a case study of precipitation in Australia from space to ground sensors

Benjamin Hines[a], Guoqi Qian*[a], Antoinette Tordesillas[a]

[a]*School of Mathematics and Statistics,*
The University of Melbourne, Parkville, 3010, VIC, Australia

Abstract

We develop a method for discovering a set of optimally representative dynamical locations (ORDL), a small subset of observed locations that are the most informative of the dynamics of a real complex system, as embodied in big spatiotemporal data. We achieve this through a two-pronged approach: (a) by reducing the multivariate time series data into a small set of time series with minimal loss of information on the dynamics of the system, (b) by exploiting the best that remote sensing and in-situ observations can offer. In the former, we extend the recently proposed empirical dynamical quantiles for univariate time series to multivariate data using a directional statistical depth measure and principal eigen-decomposition method. In the latter, we perform data fusion to leverage remotely sensed precipitation from multiple satellite platforms in addition to ground-based rain gauges to improve overall accuracy and spatial coverage. We demonstrate our method in the context of precipitation data over 2003-2021 for Australia. Of the six states, the location, ranking and number of ORDL suggest that Queensland has seen the most significant variability in precipitation while that in Victoria has remained relatively stable. Finally, this study has uncovered ungauged locations in data-sparse regions of Australia where the installation of future rain gauges can optimally represent precipitation dynamics in the region under a changing climate.

Keywords: Time Series Analysis, Multivariate Statistics, Empirical Distribution Theory, Statistical Depth, Precipitation Analysis
2000 MSC: 62M10, 62G30, 62H10

1. Introduction

Building resilience to climate change rests critically on a deep knowledge of precipitation dynamics [1]. Precipitation is a fundamental component of the Earth's water cycle which controls the distribution of water resources around the globe. The World Bank estimates that some 4 billion people are already living in water scarce areas, with two-thirds of the world's population likely to face water shortages by 2025 [2]. Water security is crucial in essentially all sectors of industry but especially in agriculture [3], energy [4], health [5], manufacturing [6] and infrastructure [7]. Extreme fluctuations in precipitation, the main trigger of a range of natural hazards like droughts [8], bushfires [9], floods [10] and landslides [11], are also seeing an increase in frequency [12]. The impacts of these can rapidly escalate and lead to cascading disasters, potentially crippling critical infrastructure and emergency services with prolonged and heavy costs to human life and property, the economy, and the environment [13].

Despite these growing global risks, progress are being made in building resilience to climate change at all levels. Broadly, this study seeks to contribute to the mitigation of these risks at the regional and, ultimately, global level, by developing tools that can efficiently harness information in big data from Earth observation systems [14]. A promising frontline defence of direct relevance to this effort is the integration of monitoring data from multiple satellite platforms with the capability for large scale monitoring — into data-driven modelling and forecasting. There are several obstacles to progress in leveraging multi-source remote sensing data for modelling and forecasting. Typically these data sets comprise big and multivariate spatiotemporal data with varying dimensions, unevenly distributed spatial locations and non-homogeneous temporal dynamics [15, 16]. Such data attributes pose serious statistical and computational challenges for analysis and modelling.

An effective strategy in overcoming these challenges, recently demonstrated in an operational landslide monitoring and early warning system [17], is dimension reduction through the concept of empirical dynamic quantiles that was originally introduced by Pena et al. [18]. This study seeks to build on these new innovations to develop a framework for discovery and application of optimally representative dynamical locations (ORDL) from big multivariate spatiotemporal data. ORDL is a small subset of observed loca-

tions that are the most informative of the dynamics of the studied system. To establish ORDL, we employ a two-pronged approach that, firstly, extends the methodology of empirical dynamical quantiles [18] to multivariate time series data by developing a multivariate ordering combining statistical depth with eigenvalue decomposition and, secondly, performs data fusion to exploit the best that in-situ and remotely sensed observations can offer for a given region.

We emphasise that our proposed method is designed to have broad applicability in the field of spatiotemporal data analytics. That said, to give a concrete application of its utility, we apply our method to a case study of precipitation data at the regional level – here the entire Australian continent. The period of study is almost two decades from 2003 to 2021 and integrates precipitation data from the rainfall gauge network of the Australian Bureau of Meteorology and two satellite platforms Global Satellite Mapping of Precipitation (GSMaP) from the Japan Aerospace Exploration Agency (JAXA) and the Climate Prediction Center Morphing Technique (CMORPH) satellite from the National Oceanic and Atmospheric Administration (NOAA).

Our aim in this case study on Australia is twofold. The first and main aim is to identify the locations in both data-rich and data-sparse areas of Australia which best represent the precipitation dynamics over the past two decades. The second is to identify parts of Australia (e.g., state level) which have endured the most significant variability under the changing climate of 2003-2021, as evident in the changing locations, ranking and number of ORDL.

Given that vast portions of Australia are currently ungauged [19], of particular interest is to uncover locations that are dynamically significant – but which remain without rain gauges. That is, we designed ORDL to potentially provide useful and actionable intelligence for policymakers and government agencies charged with water management and disaster preparedness nationwide. Clearly, the future expansion of the rain gauge network in the region will play a critical role in building resilience to climate change [20]. In fact, the optimisation of rain gauge network expansion has already been implemented in the context of landslide risk management in other parts of the world (e.g., [21]). Either through removing redundant gauge stations to reduce costs or expanding the gauge network to reduce estimation uncertainty [22, 23]. But a crucial missing element in all of these studies is consideration for the essential dynamics of precipitation and/or of landslides *under a changing climate*. ORDL offers this capability.

The paper is arranged as follows. In Section 2, we present the data for this study. Next we discuss the different components of our method in Section 3. Results and their significance are presented in Section 4. Concluding remarks are given in Section 5.

2. Australian Precipitation Data

2.1. Australian Bureau of Meteorology Rain Gauge Data

Rain gauge stations in Australia are operated by the Bureau of Meteorology (BOM). BOM is an Australian government agency responsible for providing weather research, analysis and services for Australia and the Australian regions. As Australia is a sparsely populated country, where most of the population live on the coast [24], the locations of the rain gauges must take these factors into account. There have been nearly 4000 rain gauge stations in operation at some point between 2003 and the present, for this study use 1873 (stations that have coverage for the entire period). The left plot of Figure 1 shows the 1873 locations which precipitation is measured by rain gauges for monthly intervals (showing square root precipitation measurements for January 2021). Rain gauges provide accurate measurements for the amount precipitation that has occurred. Rain gauges are limited in their ability to have high spatial coverage due to operational costs and human requirements. There is a lot of variance in the distance between rain gauge stations. Many stations have their nearest neighbour station being less than five kilometres away, whereas some have their nearest neighbour being over 200 kilometres away. The average distance between a gauge station and its nearest neighbour is about 33 kilometres. It is not feasible to have rain gauges uniformly across all of Australia, such there is an obvious desire to use satellite estimation as a substitute.

2.2. Japan Aerospace Exploration Agency Satellite Data

The Japan Aerospace Exploration Agency (JAXA) is a Japanese government agency that focuses on research, technology development and launching of satellites into orbit. The JAXA monthly satellite estimates come as a $0.1^\circ \times 0.1^\circ$ grid over Australia, This gives 69586 observed points (locations shown in the middle plot of Figure 1). An observed point's nearest neighbour is about 11 kilometres away.

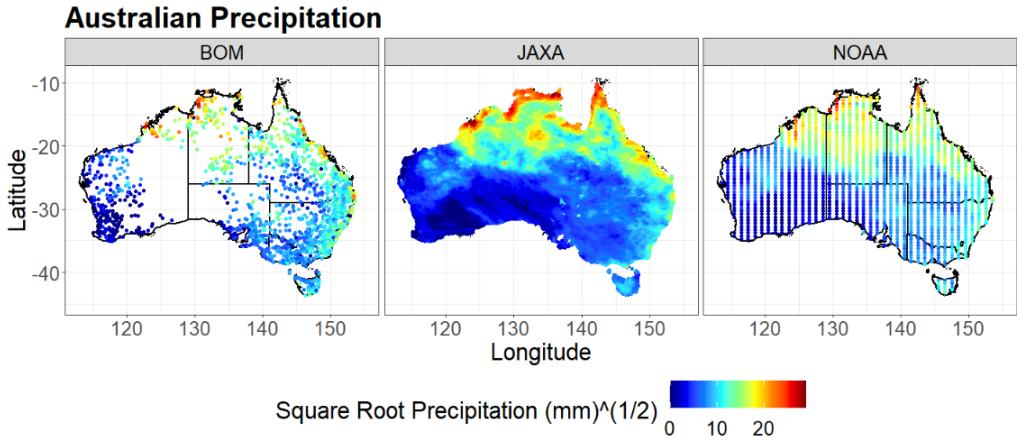


Figure 1: Locations of available precipitation data for the Bureau of Meteorology (left), Japan Aerospace Exploration Agency (middle) and National Oceanic and Atmospheric Administration (right). Showing square root precipitation measurements/estimates from January 2021.

2.3. National Oceanic and Atmospheric Administration Satellite Data

The National Oceanic and Atmospheric Administration (NOAA) is an American government agency that is responsible for many things such as weather forecasts, and oceanic and atmospheric condition monitoring. The NOAA monthly satellite estimates come as a $1^\circ \times 1^\circ$ grid over Australia. This gives 1391 observed points (locations shown in the right plot of Figure 1). An observed point's nearest neighbour is about 110 kilometres away.

3. Methodology

A summary of the method for finding the optimally representative dynamical locations for multivariate data is shown in Figure 2. In a univariate setting, empirical dynamical quantiles (EDQ) can be used to find ORDL for high-dimensional spatiotemporal data. Consider $\mathbf{x}_i' = (x_{i,1}, \ldots, x_{i,T})$ to be the spatiotemporal data at location $i = 1, \ldots, n$, at observed time intervals $t = 1, \ldots, T$. The level-p EDQ, $q_t^{(p)}$, from the observed data $\mathcal{C}_n = \{\mathbf{x}_i | i = 1, \ldots, n\}$ is derived in [18] as one of the n univariate time series satisfying

$$q_t^{(p)} = \underset{y_t \in \mathcal{C}_n}{\arg\min} \left\{ \sum_{t=1}^{T} \left(\sum_{x_{i,t} \geq y_t} p|x_{i,t} - y_t| + \sum_{x_{i,t} < y_t} (1-p)|x_{i,t} - y_t| \right) \right\}. \quad (1)$$

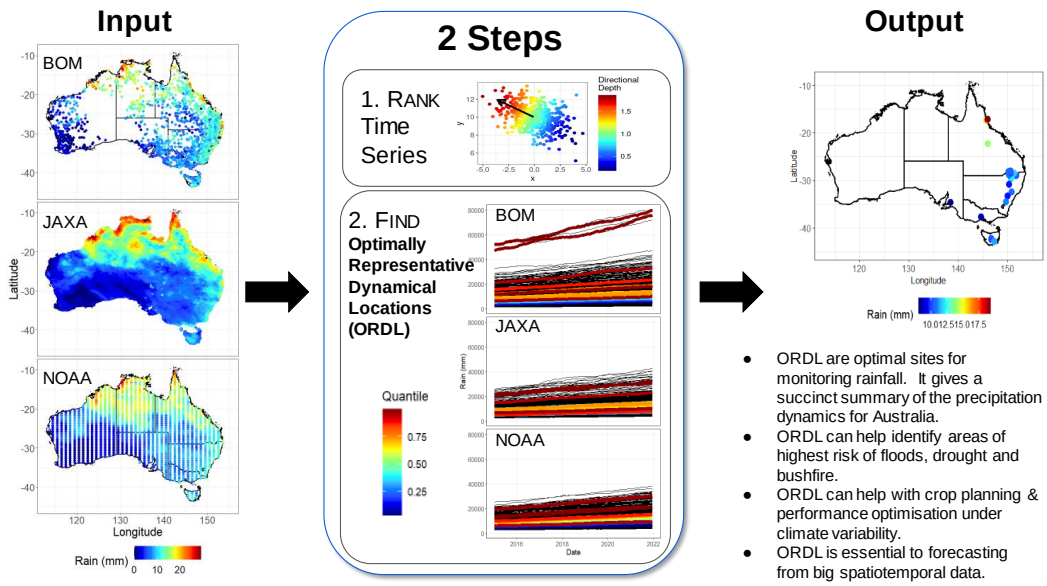


Figure 2: Flowchart demonstrating the process of finding a set of optimally representative dynamical locations for precipitation in Australia through the multivariate empirical dynamical quantiles.

An empirical dynamical quantile is considered an observed time series which describes the evolution of the temporal data for some level of descriptive quantile. Empirical dynamical quantiles not only allow us to find the ORDL for univariate data, they also allows us to assign it a numeric value based on what section of the data set it represents. To find the ORDL for a multivariate spatiotemporal data set, we cannot simply use empirical dynamical quantiles. EDQ in the current form, cannot be used on multivariate data. The formula in equation (1) must be altered to allow for multivariate empirical dynamical quantiles to be used to find the multivariate ORDL. The process of finding the ORDL of multivariate data is divided into two phases. Firstly, give a multivariate ordering of the time series at each observed time interval. Secondly, calculate the multivariate empirical dynamical quantiles.

3.1. Step 1: Rank Time Series

A summary for ranking the multivariate time series data is demonstrated in Figure 3. In equation (1) the sum functions are dependent on the ordering of the univariate data. With multivariate data, ranking the observations through some ordering is not as simple. We develop a process based on sta-

tistical depth to rank multivariate observations. This is a two step process, where firstly we use statistical depth a measure to give the multivariate data a quantile like centre outward description. Then using eigenvalue decomposition, give direction to the resulting values from the previous step to allow for a linear ordering.

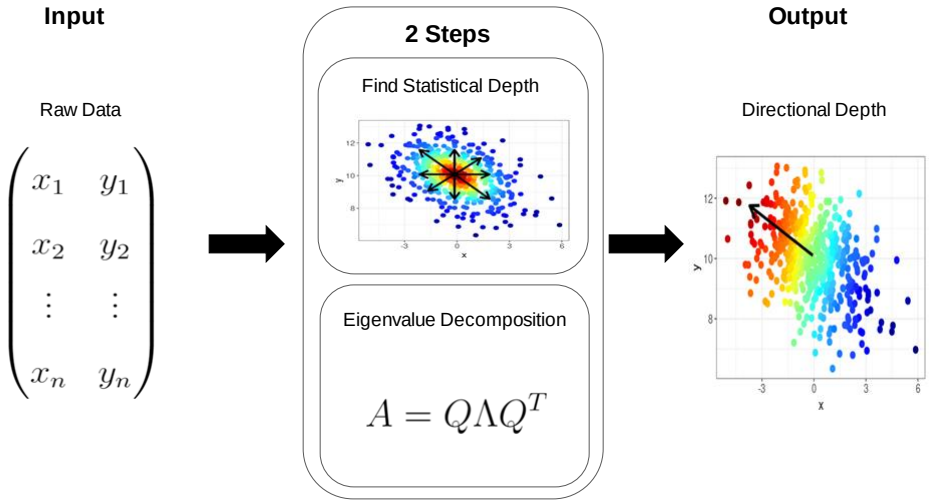


Figure 3: Flowchart demonstrating the 2 step process for obtaining a multivariate quantile like ordering.

For univariate data, quantiles can be used to give a simple ranking of the values in a data set [25]. Quantiles for multivariate data cannot be calculated or interpreted in the same way. Consider n $d-$dimensional random samples $\mathbf{x}_i \in \mathbb{R}^d$ for $i = 1, \ldots n$ from a multivariate random variable. For a sample of multivariate data, we can use what is known as statistical depth to give a quantile like description of the data set [26]. Statistical depth can be described as a median oriented quantile function, where multivariate points are assigned a value based on their relative distance from the multivariate median of the data. There are several different methods for calculating the depth of a point in a data set. For example, the Mahalanobis depth of a point $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^T \in \mathcal{S}_n \subset \mathbb{R}^d$ is defined as

$$D(\mathbf{x}_i; \mathcal{S}_n) = \left[1 + (\mathbf{x}_i - \bar{\mathbf{x}})^T S^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})\right]^{-1}$$

where $\mathcal{S}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is the sample of data points, $\bar{\mathbf{x}}$ is sample mean

vector of $\mathcal{S}_n$ and S is the $d \times d$ sample covariance matrix of $\mathcal{S}_n$ [27]. The depth of a point relative to a given data set gives a measurement of how deep that point lies in the data cloud [28]. There are several other depth functions which can be used [29, 30].

Figure 4 shows positively correlated 2-dimensional data simulated from a bivariate normal distribution. The blue triangle represents the Mahalanobis deepest (largest depth) point in the data set. This point is roughly in the centre of the data cloud. Statistical depth has been used in several areas of study, such as economics [31] and spatial statistics [32]. The properties that

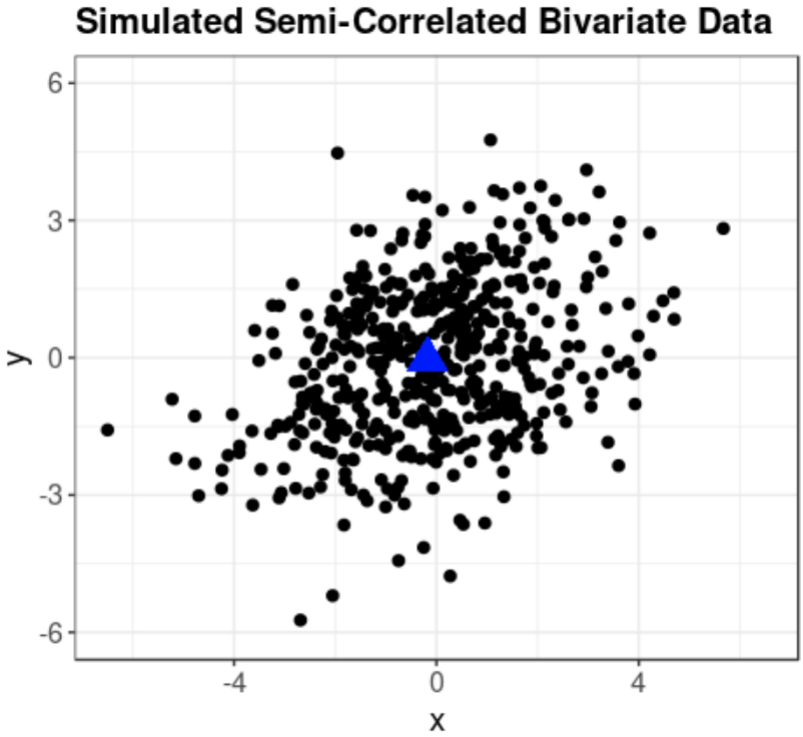


Figure 4: Simulated positively correlated bivariate normal data. Blue triangle represents the point with the largest (deepest) Mahalanobis depth.

a well defined depth function are expected to have has been discussed in [30].

While depth can give us a quantile like description of multivariate data, it is not entirely comparable to a quantile function. The centre outward ordering of the data means the depth function describes how far points are away from centre of the data cloud. The absence of a directional component

results in the points in a data set not able to be linearly ordered. This can be shown in a univariate setting, consider the data points of $1, 2$ and 3, the depth of these points are $\frac{1}{3}, \frac{2}{3}$ and $\frac{1}{3}$ respectively. While there is an order to this univariate data, depth cannot give an entirely comparable quantile description. We need to make an adjustment to the notion of depth, in order for it to have a directional aspect, such that it is more comparable to a quantile function.

Define $\mathbf{x}^{(m)}$ to be the deepest point with respect to the data set $\mathcal{S}_n$ for some depth function $D(\mathbf{x}_i; \mathcal{S}_n)$. If we consider the eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_d$ of the sample covariance matrix S, then the principal eigenvector $\mathbf{e}_1$ (the eigenvector corresponding to the largest eigenvalue) can give the depth function a sense of direction. If we add the principal eigenvector to the deepest point, we obtain another vector from the deepest point still in the direction of the principal eigenvector $\vec{\mathbf{v}} := \mathbf{x}^{(m)} \to \mathbf{x}^{(m)} + \mathbf{e}_1$. Let $\vec{\mathbf{u}}_i$ be the vector from the deepest point, to the point $\mathbf{x}_i$ ($\vec{\mathbf{u}}_i := \mathbf{x}^{(m)} \to \mathbf{x}_i$). If $\vec{\mathbf{u}}_i$ has a positive projection onto $\vec{\mathbf{v}}$, then consider the point $\mathbf{x}_i$ to have a larger directional depth than $\mathbf{x}^{(m)}$. The sign of the projection of $\vec{\mathbf{u}}_i$ onto $\vec{\mathbf{v}}$ can be expressed in terms of the angle between these vectors. If we let θ_i be the angle between $\vec{\mathbf{u}}_i$ and $\vec{\mathbf{v}}$, then the projection of $\vec{\mathbf{u}}_i$ onto $\vec{\mathbf{v}}$ is positive if $0 \leq \theta_i \leq \frac{\pi}{2}$, and negative if $\frac{\pi}{2} < \theta_i \leq \pi$, where $\theta_i = \arccos \frac{\vec{\mathbf{u}}_i \cdot \vec{\mathbf{v}}}{||\vec{\mathbf{u}}_i||_2 ||\vec{\mathbf{v}}||_2}$ and $||\cdot||_2$ is the ℓ_2-norm. This new depth of the point $\mathbf{x}_i$ is expressed as

$$D'(\mathbf{x}_i; \mathcal{S}_n) = \begin{cases} D\big(\mathbf{x}^{(m)}; \mathcal{S}_n\big) + \big(D\big(\mathbf{x}^{(m)}; \mathcal{S}_n\big) - D(\mathbf{x}_i; \mathcal{S}_n)\big) & , \text{ if } 0 \leq \theta_i \leq \frac{\pi}{2} \\ D(\mathbf{x}_i; \mathcal{S}_n) & , \text{ otherwise} \end{cases} . \tag{2}$$

Figure 5 shows an example of this for a bivariate case. The arrow from the deepest point indicates the direction of the principal eigenvector. The green points have a positive projection onto $\vec{\mathbf{v}}$ and therefore are considered to have a larger directional depth than $\mathbf{x}^{(m)}$. Whereas the red points have a negative projection onto $\vec{\mathbf{v}}$ and are therefore considered to have a smaller directional depth than $\mathbf{x}^{(m)}$. There are times where the principal eigenvector direction may be negative in one or more dimensions. In a bivariate case, this can occur when the two variables are negatively correlated and thus the principal eigenvector will be negative in one dimension and positive in the other. As we are trying to give direction to multi-dimensional data, this will not be a problem as it will still help describe the underlying trend. Equation (2) is not the only way the directional depth can be defined. We can define a

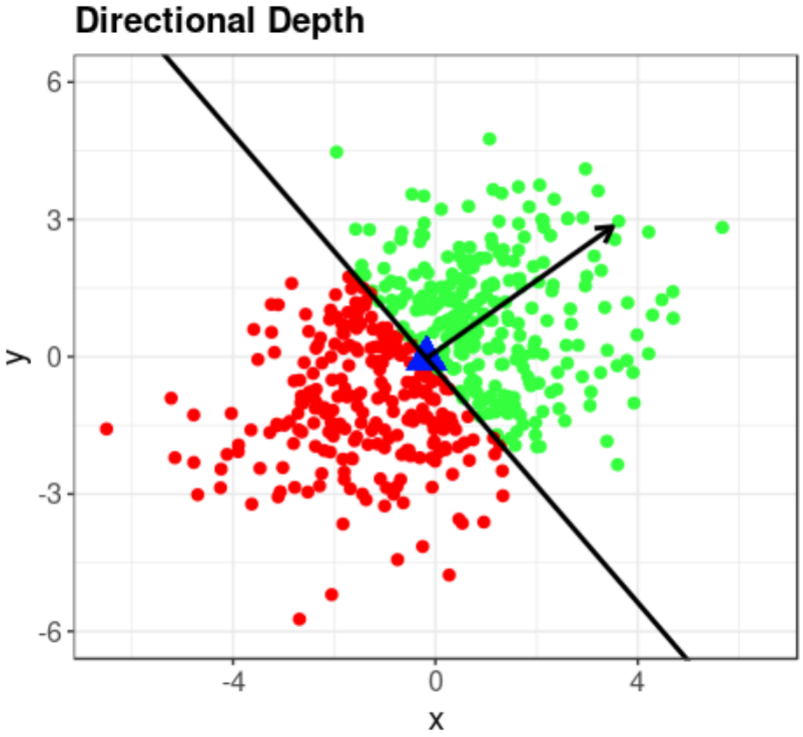


Figure 5: Directional depth sign of simulated positively correlated bivariate normal data. Green points have a positive directional depth (positive projection onto principal eigenvector) and red points have a negative directional depth (negative projection onto principal eigenvector).

directional depth function which instead of being a piecewise function, uses a weighting which depends on the angle. I.e.

$$D'(\mathbf{x}_i; \mathcal{S}_n) = \left(1 - \frac{\theta_i}{\pi}\right)\left(2D\big(\mathbf{x}^{(m)}; \mathcal{S}_n\big) - D(\mathbf{x}_i; \mathcal{S}_n)\right) + \frac{\theta_i}{\pi} D(\mathbf{x}_i; \mathcal{S}_n)\,. \quad (3)$$

The difference between the piecewise and weighted directional depth for negatively correlated bivariate data is shown in Figure 6. The weighted directional depth gives a more gradual transition of value, whereas the piecewise directional depth has a very clear point between positive and negative. This is very evident for points, in which the angle between the two vectors is close to $\frac{\pi}{2}$. The choice of directional depth function will have a more significant affect for the points that are closer to deepest point, and a less significant

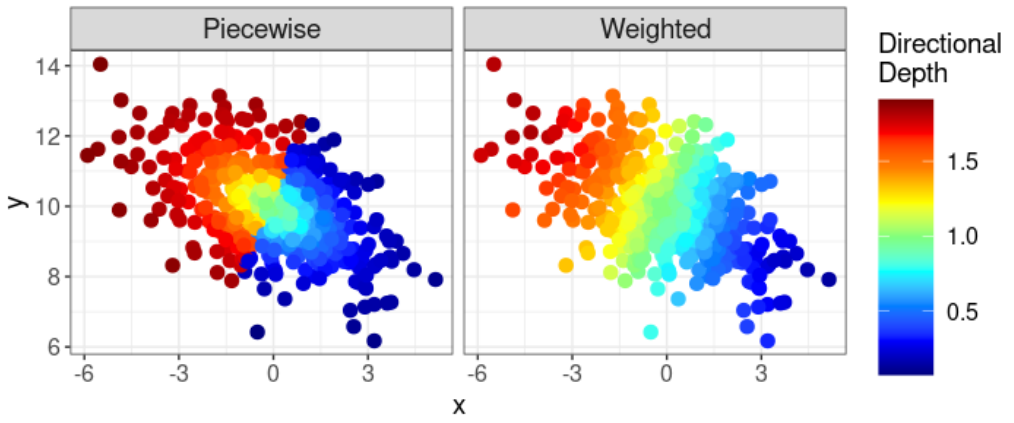


Figure 6: Directional depth for negatively correlated bivariate data using the piecewise method (left) and the weighted method (right) with the direction of the principal eigenvector from the deepest point.

affect to the points which are far away from the deepest point, and have an angle close to 0 or π.

3.2. Step 2: Finding Multivariate Empirical Dynamical Quantiles

The univariate EDQ formula in equation (1) uses the ordering of the data to give each time series a ranking at each interval. Instead we use the directional depth function as in equations (2) or (3) for this. We also replace the absolute difference with an ℓ_2-norm function (note that other distance metrics could be used). Then a level-p ($p \in [0,1]$) multivariate empirical dynamical quantile (MEDQ) can be defined as following

$$\mathbf{q}_t^{(p)} = \underset{\mathbf{y}_t \in \mathcal{S}_n}{\arg\min} \left\{ \sum_{t=1}^{T} \left(\sum_{D'(\mathbf{x}_{i,t};\mathcal{S}_{n,t}) \geq D'(\mathbf{y}_t;\mathcal{S}_{n,t})} p||\mathbf{x}_{i,t} - \mathbf{y}_t||_2 + \sum_{D'(\mathbf{x}_{i,t};\mathcal{S}_{n,t}) < D'(\mathbf{y}_t;\mathcal{S}_{n,t})} (1-p)||\mathbf{x}_{i,t} - \mathbf{y}_t||_2 \right) \right\} \tag{4}$$

where $\mathbf{x}_{i,t}$ is the i^{th} multi-dimensional vector at time t, $\mathcal{S}_{n,t}$ is the collection of $\mathbf{x}_{i,t}$'s at time t with $i = 1, \ldots, n$ and $||\cdot||_2$ is the ℓ_2 norm. The directional depth function is calculated relative to the data at each time interval. First the depth of each time series is calculated for each interval. Then the directional component is included via the eigenvalue decomposition. It should be noted that similarly in the univariate case, different values of p, may not

give different time series as the multivariate empirical dynamical quantiles. The same time series may minimise the function in equation (4) for a large range of p. A multivariate empirical dynamical quantile can be described similarly to the univariate empirical dynamical quantiles, with the addition it now describes the evolution of temporal multivariate data, for some level of descriptive quantile, across all variables.

As we often will want a specific number of ORDL for dimension reduction, we apply the function in equation (4) for a set of quantiles $P = \{p_1, \ldots, p_k\}$. The number of unique time series chosen by multivariate empirical dynamical quantiles is a function of two inputs. Firstly, the size of the multivariate time series and their overall consistency in behaviour. By consistency, we mean a given time series for each interval, will maintain relatively the same directional depth ranking. Secondly, the size and nature of the set P. The consistency of the multivariate time series is the largest factor in determining the number of unique MEDQs will be obtained. If a time series is not consistent, it is less likely to be able to minimise the function in equation (4) as the directional depth ranking will change at different intervals. If many of the multivariate time series are inconsistent, then fewer will be able to minimise the function and thus we will have a smaller set of representative time series for the same set P.

In obtaining ω ORDL, it is desired they share the same nature of diversity as the multivariate data set, as to give proper representation. To ensure this we choose the set P such that elements are uniformly selected between 0 and 1. For example, if we wanted to obtain ω ORDL, we would initially look at the multivariate empirical dynamical quantiles for $P = \left\{0, \frac{1}{\omega-1}, \frac{2}{\omega-1}, \ldots, 1\right\}$. Note this may not give *omega* ORDL, as the same multivariate time series could minimise the function equation (4) for multiple values in the set P. Thus, certain changes may need to be made to the elements in P to ensure we obtain *omega* ORDL.

3.3. Simulations

We will show two simulated examples for different situations, to demonstrate the ability of multivariate empirical dynamical quantiles to give representative observations in multivariate temporal data sets. In the first example, we will look to find the multivariate empirical dynamical quantile of bivariate time series data where the variables are positively correlated, for $P\{0.05, 0.5, 0.95\}$. Each time series is a simulated stationary autoregressive model of order 1 ($AR(1)$) [33], where the i^{th} time series in each set share

a common initial value relative to the mean of the set with some error. In other words,

$$\begin{aligned} x_{i,t}^{(1)} &= \phi_1 x_{i,t-1}^{(1)} + \varepsilon_{i,t}^{(1)} \\ x_{i,t}^{(2)} &= \phi_1 x_{i,t-1}^{(2)} + \varepsilon_{i,t}^{(2)} \end{aligned} \tag{5}$$

where $\phi_1 = 0.8897$, $\varepsilon_{i,t}^{(1)} \overset{i.i.d.}{=} \varepsilon_{i,t}^{(2)} \sim \mathcal{N}(0, 0.4)$, $x_{i,1}^{(1)} \sim Z_i + \varepsilon_{i,1}^{(1)}$, $x_{i,1}^{(2)} \sim Z_i + 10 + \varepsilon_{i,1}^{(2)}$ and $Z_i \sim \mathcal{N}(0,1)$ independent of $\varepsilon_{i,t}^{(1)}$ and $\varepsilon_{i,t}^{(2)}$. These two sets of realisations are positively correlated, in that they have initial values that are up to some error, the same relative to the mean of the entire set, and as they are stationary they should have consistent directional depth rankings. Figure 7 shows the multivariate EDQ for the level-0.05 MEDQ (red), level-0.5 MEDQ (green) and level-0.95 MEDQ (blue) quantile time series. The MEDQ time series given exhibit strong positive correlation in their re-

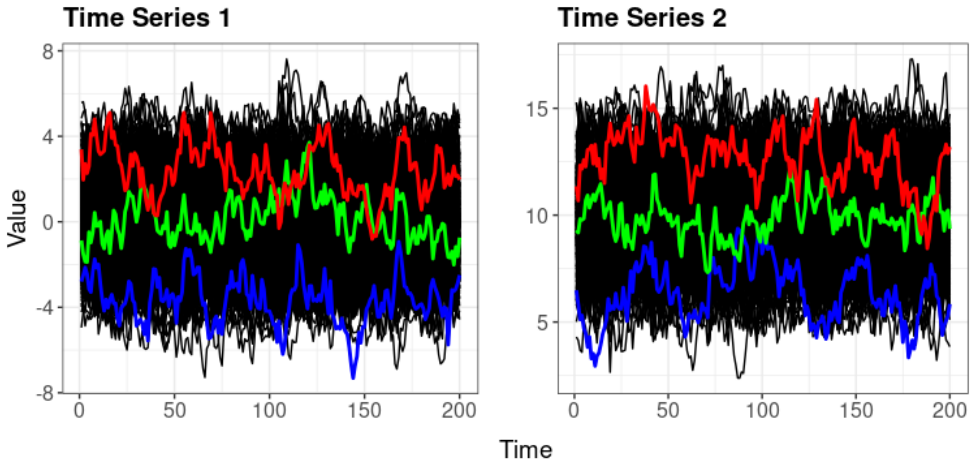


Figure 7: Three representative multivariate time series from set of 500 positively correlated time series. Representative set chosen by the multivariate empirical dynamical quantiles for $p = 0.05$ (blue), $p = 0.5$ (green), $p = 0.95$ (red) using the piecewise directional depth function.

spective position in their data sets. While there is some cross over between the representative time series, they still represent the data at each of their respective quantiles and are located in reasonable positions relative to the data set,

Another example of multivariate empirical dynamical can come from the

situation where four temporal variables, which are both positively and negatively correlated. That is,

$$\begin{aligned} x_{i,t}^{(1)} &= \phi_1 x_{i,t-1}^{(1)} + \varepsilon_{i,t}^{(1)} \\ x_{i,t}^{(2)} &= \phi_1 x_{i,t-1}^{(2)} + \varepsilon_{i,t}^{(2)} \\ x_{i,t}^{(3)} &= \phi_1 x_{i,t-1}^{(3)} + \varepsilon_{i,t}^{(3)} \\ x_{i,t}^{(4)} &= \phi_1 x_{i,t-1}^{(4)} + \varepsilon_{i,t}^{(4)} \end{aligned} \tag{6}$$

where $x_{i,1}^{(1)} \sim Z_i + \varepsilon_{i,1}^{(1)}$, $x_{i,1}^{(2)} \sim -2Z_i + 10 + \varepsilon_{i,1}^{(2)}$, $x_{i,1}^{(3)} \sim Z_i - 5 + \varepsilon_{i,1}^{(3)}$ and $x_{i,1}^{(4)} \sim -Z_i - 10 + \varepsilon_{i,1}^{(4)}$ (ϕ_1, Z_i and $\varepsilon_{i,t}^{(1,2,3,4)}$ are defined the same as above).

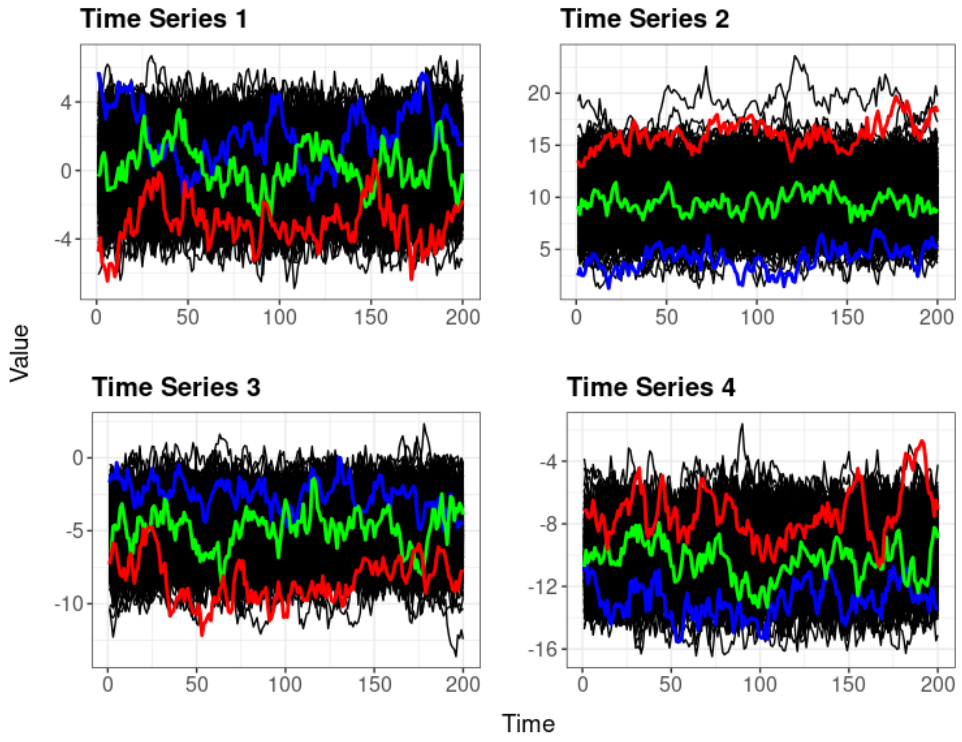


Figure 8: Three representative multivariate time series from set of 500 positively and negatively times series. Representative set chosen by the multivariate empirical dynamical quantiles for $p = 0.05$ (blue), $p = 0.5$ (green), $p = 0.95$ (red) using the weighted directional depth function.

Figure 8 shows the multivariate empirical dynamical quantiles for 4-

dimensional simulated temporal data. The top left plot shows the time series corresponding to $\mathbf{x}^{(1)}$, which has positively correlated initial values to the time series in the bottom left plot ($\mathbf{x}^{(3)}$), but negatively correlated with the time series in the top right ($\mathbf{x}^{(2)}$) and bottom right ($\mathbf{x}^{(4)}$). Having negatively correlated initial values does not affect the overall ability to obtain representative time series. The level-0.05 MEDQ (blue) is in the upper parts of the top left and bottom left plots, but in the lower parts of the top right and top left plots. Similarly the level-0.95 MEDQ (red) is in the lower parts of the top left and bottom left plots, but in the upper parts of the top right and bottom right plots.

There are situations where multivariate empirical dynamical quantiles will not be able to give optimally representative dynamical locations. If the variables in the data set are uncorrelated, the directional depth ordering at each time interval will be inconsistent, and thus it is unlikely that a multivariate time series would be able to give an informative representation of the trend of the multivariate data. There are also situations where alterations may need to be may to the data. If the relative range of the variables is significantly different, the variable with the significantly larger range will influence the ℓ_2-norm in equation (4) more than the others. Consequently, the multivariate time series that minimises the function will be the time series that minimises the variable with the largest range. Similarly to techniques like principal component analysis, to obtain results that are most informative, standardisation of the data should be considered [34].

3.4. Implementation for Australian Precipitation Data

Precipitation is a spatially dependent seasonal environmental process. Therefore locations will receive their annual highs and annual lows at different points in the year. While Australia, for the most part, has long-term stationary precipitation trends [35], the seasonality of precipitation means that the short-term trends are non-stationary. Consequently, the directional depth ranking will be very variable across the time intervals and therefore majority of the time series will not be consistent. Thus in raw form, the optimally representative dynamical locations would not give us a lot of information. To overcome the seasonality of this data, we use a cumulative sum of precipitation for each location. This would allow for each time series to be more distinguishable, and therefore the directional depth ordering at each time interval would be more consistent. This is not an uncommon practice for precipitation. For example, the standardised precipitation index requires

the cumulative precipitation amount for a given period of time, of a given location, to give a measurement of a meteorological drought [36, 37].

3.4.1. Gauge and Satellite ORDL Comparison

Empirical dynamical quantiles will be used to obtain 20 ORDL for the BOM rain gauge precipitation data. This will be compared to the 20 ORDL for the JAXA and NOAA satellite precipitation data, found through the multivariate empirical dynamical quantiles using the same spatial domain as the gauge data set ($n = 1873$). The satellite estimates are placed on grid points over Australia, thus the observed points are not at the exact same locations as rain gauge stations. We perform a simple inverse distance weighting to obtain interpolated values for the satellite estimates at the rain gauge locations [38].

3.4.2. Complete Satellite ORDL

Multivariate empirical dynamical quantiles will be used to obtain 20 ORDL for the JAXA and NOAA satellite estimates, on their entire spatial domain. The NOAA observed locations are a subset of JAXA, thus we use only the 1391 NOAA locations for this calculation. Note that there are no gauge measurements for 372 of these locations.

3.4.3. BOM-Satellite and Fused-Satellite ORDL Comparison

The aim of fusing gauge and satellite precipitation data is to obtain accurate precipitation with high spatial coverage. There have been many studies done on fusing gauge and satellite precipitation data through many different methods [39, 40, 41, 42]. As fusing gauge and satellite precipitation data is not the focus of this paper, we will not go into significant detail about the methodology. We will be extending and modifying the two-step process outlined in [43] to multiple satellite sources. Firstly we express the relationship between the square root of the gauge measurements and the square root of the satellite estimates at each location as a time invariant linear equation. Then perform spatial interpolation on the regression coefficients using a geographically weighted model to obtain coefficient estimates in locations with no rain gauges. This then allows us to obtain fused precipitation estimates in all parts of the country, for every time interval. The ultimate goal of fusing gauge and satellite precipitation data is to have a single variable which can accurately represent the true nature of precipitation with high spatial coverage. However we cannot simply disregard the data used to obtain these

fused values. These fused data points will always be dependent on the satellite estimates. Thus we consider the fused and satellite data together as a multivariate data set with high coverage of locations in Australia. We will compare 20 ORDL from the fused and satellite data on the 2245 locations (gauge and satellite locations), to 20 ORDL from the joint BOM and satellite multivariate data on the 1873 locations.

3.5. Fused-Satellite ORDL evolution

The number of ORDL is simply a product of the input probability set P of quantiles to find. Through changing the values in the set P, we may change the number of optimally representative dynamical locations that we obtain. Note that as we increase the number of ORDL, ω, the output may not always be the same. This is due the ways we increase the size of the set to give ORDL. As we want the ORDL to truly represent the underlying dynamics of the system, we use values of p between 0 and 1 that are equally spaced. For example, if we want a collection of 10 ORDL, we may use the set $P_{10} = \left\{0, \frac{1}{9}, \ldots, 1\right\}$. However, if we want a collection of 20 representative time series, we may use the set $P_{20} = \left\{0, \frac{1}{19}, \ldots, 1\right\}$. As the values in each of these sets are different (excluding 0 and 1), a time series that minimise the function in equation (4) for values in P_{10}, may not minimise the function for values in P_{20}. Intuitively, this makes sense as if we increase the number of representative time series, each representative can represent a fewer number of observations and thus can have more specific behaviour. For a simple example, consider the set of $A = \{1, 2, 3, 4, 5\}$. If we wanted a single number to represent this set, we would choose 3. However, if we wanted two numbers to represent the set, we would choose 2 and 4. Changing the number of ORDL will show the evolution of where the representatives are located, when the amount of data they are representing changes. We will look at this evolution of the ORDL for the fused and satellite data set.

4. Results and Discussion

4.1. Gauge and Satellite ORDL Comparison

Figure 9 shows a set of 20 optimally representative dynamical locations for the BOM rain gauges (through empirical dynamical quantiles), and for the JAXA and NOAA satellite estimates (through multivariate empirical dynamical quantiles). The univariate/multivariate empirical dynamical quantiles are found for $P = \left\{0, \frac{1}{19}, \ldots, 1\right\}$ using the piecewise directional depth

function in equation (3). While we know that satellite estimates are not necessarily accurate, they can provide an indication for the amount of precipitation that has occurred [44]. It is also known both the JAXA and NOAA satellite estimates tend not to perform as well for very small or very large amounts of precipitation, over and underestimating what has been recorded respectively [45]. These results confirm these past studies, as the locations and values of the representative time series vary significantly due to the differences in the underlying dynamics of each system. Specifically for the BOM EDQ results, the largest quantile (or time series representing the locations that obtain the most cumulative precipitation) is in north eastern Australia, the wettest part of the country [46]. However, the satellite MEDQ does not have a representative in that region, potentially due to the fact that satellites underestimate very high levels of precipitation. Note that if we were to increase the number of ORDL, we could potentially see the satellite MEDQ give a representative in north eastern Australia.

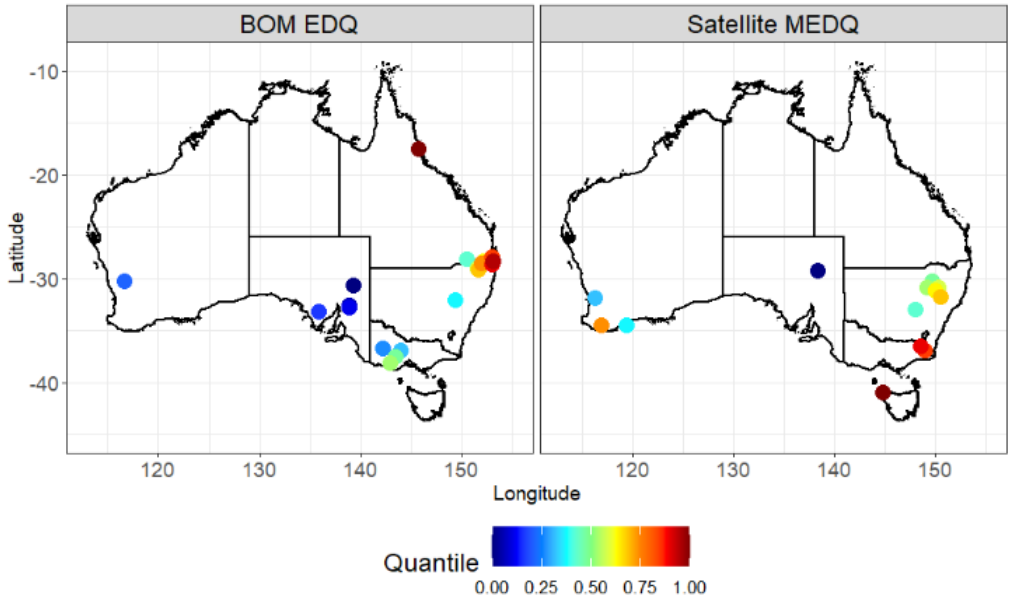


Figure 9: Reduction in dimension to 20 representative time series' locations for the univariate cumulative precipitation of BOM rain gauges (left, n = 1873), and the multivariate cumulative precipitation of JAXA and NOAA satellite estimates (right, n = 1873). Representatives chosen through univariate/multivariate empirical dynamical quantiles.

These results show the underlying dynamics of the systems are different,

however it needs to be noted satellite precipitation estimation is a constantly improving method [47, 48]. Thus we may expect in the future, these ORDL converge to similar locations.

While the ORDL we have found are locations based on the sample of either, the true gauge data or the satellite data, we do not know if the give truly representative locations of the true precipitation system. They are however consistent with past studies on precipitation in Australia. As we have been looking at cumulative precipitation, and for the most part, long-term precipitation in Australia is stationary [35], the values of the ORDL should coincide with the quantiles of annual precipitation. Even though the results show the ORDL are different for both data sets, the locations of the representatives are consistent with the spatial distribution of annual precipitation in Australia [49, 50].

4.2. Complete Satellite ORDL

Figure 10 shows the 20 ORDL for the JAXA and NOAA joint data sets across the 1391 available locations. This is contrasting to the results in Figure 9. The ORDL are now more spread out across the country. There are several multivariate empirical dynamical quantiles in locations where there are no rain gauges available. Again, we know these satellite estimates are not completely accurate, but they give an indication of how much precipitation has occurred, and are considered as useful sources for precipitation estimation over Australia [45]. By obtaining representatives in regions where there are no rain gauges available, indicates we may be missing out on important information if we were to model precipitation using just rain gauge information. This reinforces the desire to fuse the gauge and satellite data, as to obtain values of precipitation that are both accurate, and also have a high spatial coverage.

In finding the ORDL for the multivariate satellite data over the entire domain as in Figure 10, not only shows us that there may be important information lost in just using gauge stations, it can also inform future policy. These ORDL can show us where to optimally place any future rain gauge stations such that they have the most significant impact in helping to validate and calibrate satellite estimates. As there is a financial and human requirement in operating rain gauge stations, it is important that any new rain gauge station be placed in a way such that it provides optimal additional precipitation information [51].

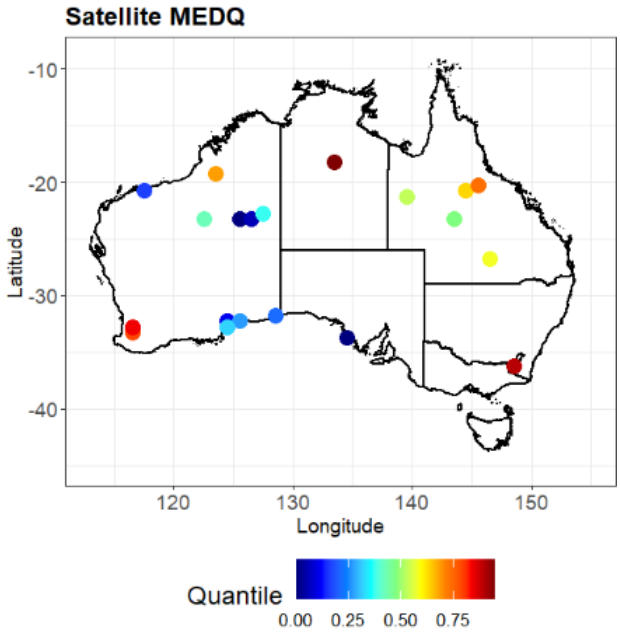


Figure 10: Reduction in dimension to 20 representative time series' locations for the multivariate cumulative precipitation of JAXA and NOAA satellite estimates (n = 1391). Representatives chosen through multivariate empirical dynamical quantiles.

4.3. BOM-Satellite and Fused-Satellite ORDL Comparison

Figure 11 shows us the locations for 20 representative multivariate time series for the BOM and satellite multivariate data, and the fused and satellite multivariate data. For the BOM and satellite multivariate data, we find the 20 representatives from the 1873 locations where rain gauges are available. For the Fused and Satellite multivariate data, we find the 20 representatives from the 2245 locations (gauge locations and satellite locations with no rain gauges). There are significant differences between the ORDL of the two data sets. The representatives found using the BOM and satellite data are spread around the south and east coasts, where many rain gauges are found. The representatives found using the fused and satellite data are more evenly distributed around the country. The fused and satellite MEDQ also give significantly different results to what the pure satellite MEDQ gave in Figure 10. By fusing the gauge and satellite data, we are possibly better able capture dynamics in the spatial variation of precipitation, that satellite my not be able to capture. For heavy precipitation satellites tend to underestimate the

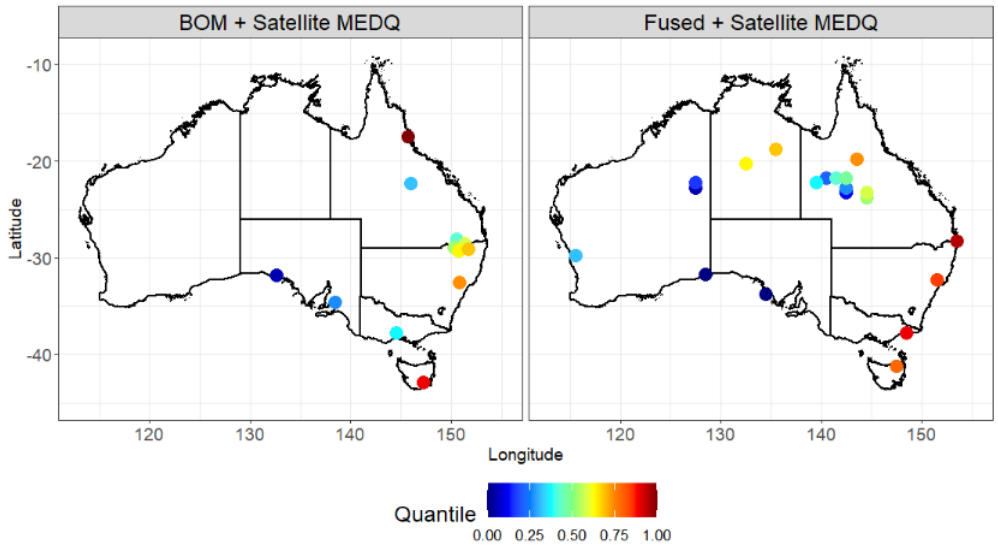


Figure 11: Reduction in dimension to 20 representative time series' locations for the multivariate cumulative precipitation of BOM rain gauges and the satellite estimates (left, n = 1873), and the multivariate cumulative precipitation of fused and satellite estimates (right, n = 2245). Representatives chosen through multivariate empirical dynamical quantiles.

true value [45], this is a problem due to heavy precipitation events having short-range spatial dependencies [52]. Thus the satellite ORDL may not be able to properly represent this climate variability in locations with high frequency of heavy precipitation events. The fused and satellite ORDL may be able to truly represent this climate variability.

These ORDL can also provide information for policy regarding precipitation extreme events. In the right plot of Figure 11, the highest quantile location is on the east coast near the city of Lismore. In late February and early March of 2022, significantly increased amounts of rain occurred over the east coast of Australia causing significant flooding. In the space of 2 days, Lismore received over 550mm of precipitation (sourced from the BOM). These ORDL may be used as a guide to determine where best to implement flood prevention measures. Optimal placement of flood mitigation measures is likely to reduce the loss of life and property damage in future flooding events [53, 54].

4.4. Fused-Satellite ORDL evolution

Figure 12 shows how these ORDL for the fused and satellite multivariate data change as we increase the size of the set P for the multivariate empirical dynamical quantiles. Some locations once they are included as a representative dynamical locations, they are consistently present. However, there are some locations where this is not the case. For example, for the 10 representatives set, the point representing the $p = \frac{6}{9}$ multivariate empirical dynamical quantile is located in north western Australia. This location is not an element of 20 representative set. As we increase the number of ORDL, we would expect to see fewer locations being removed from the representative set. This is due to there being less difference of the values in the set P, for larger values of ω, when using equally spaced intervals. In Figure 12, as the number of representatives increase, we obtain a more spatially diverse set of locations. However, a lot of the new representatives can be in locations near already present representatives for a smaller set. This raises a question about what the optimal number of locations you would need to best represent the true underlying dynamics of the multivariate system, while still reducing the dimensions of the data set to a significant degree. For a large number of representative locations, the fused and satellite ORDL in Figure 12, contained locations that were for the mos part similar to the BOM ORDL shown in Figure 9. This shows when we require a smaller set of ORDL, the differences in the underlying dynamics of the systems (both in accuracy and spatial resolution) prevent their respective representative locations from being similar. However, when the number of ORDL increases, we may be able to capture some similar representative locations, due to the correlated nature between gauge measurements and satellite estimates.

Similarly to the results in Figure 11, these results can be used to inform future policy about rain gauge station construction. The important distinction between the results is the fused and satellite ORDL give information about the best locations for rain gauge placement for the process of precipitation modelling rather than satellite validation and calibration.

Also shown in Figure 12 is the number of ORDL in each state/territory for each set of representatives. The number of ORDL in each state/territory is not necessarily proportional to the size of the region. The state of Queensland (QLD), accounts for roughly 22.5% of Australia's total land area. For a set of 10 ORDL, QLD accounts for 30% of the representatives. However, for a set of 60 ORDL, this increases to around 37%. The state of Victoria (VIC) on the other hand, which accounts for only 3% of Australia's total

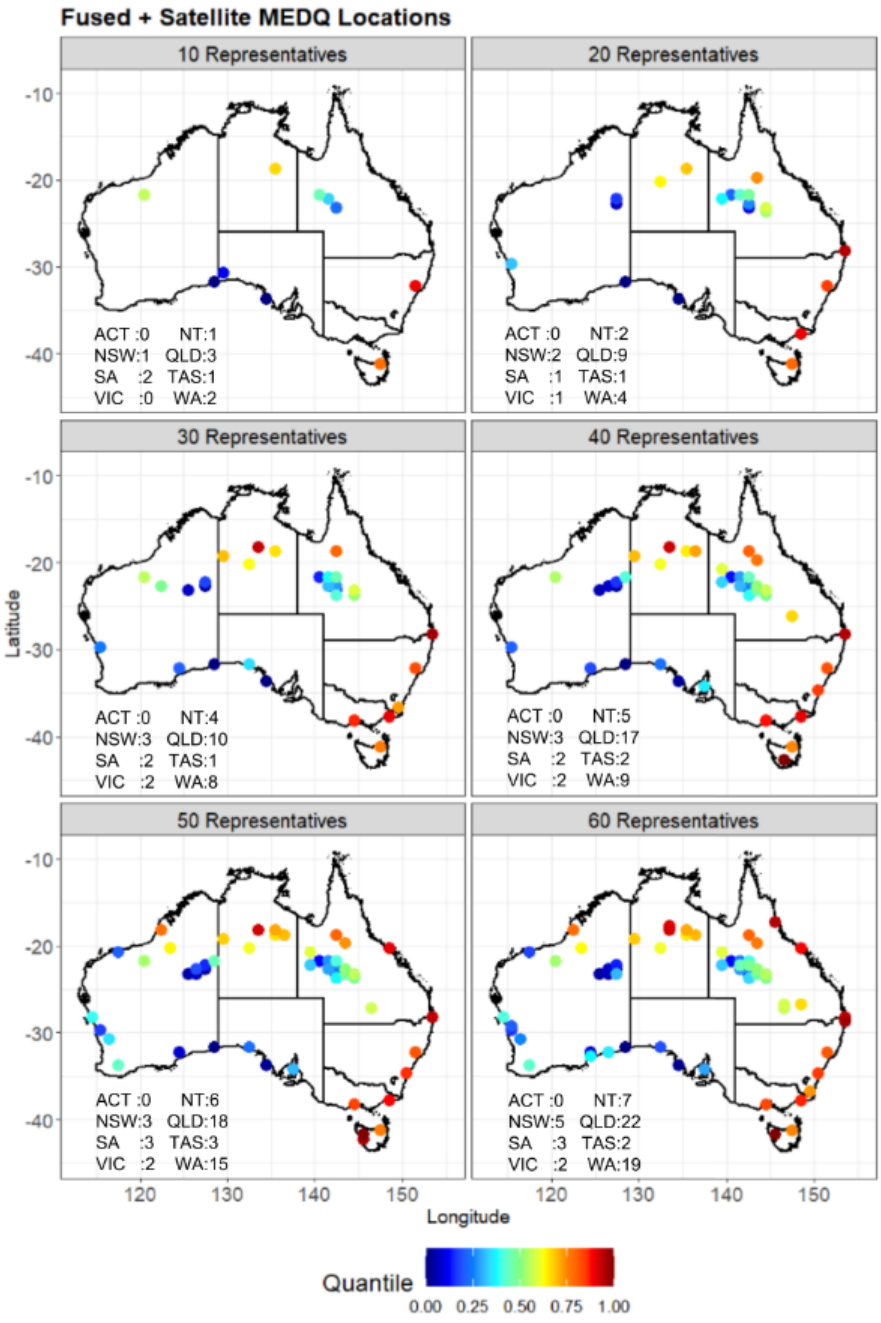


Figure 12: Reduction in dimension to 10 (top left), 20 (top right), 30 (middle left), 40 (middle right), 50 (bottom left) and 60 (bottom right) representative time series' locations for the multivariate cumulative precipitation of JAXA and NOAA satellite estimates (right, $n = 2245$). Representatives chosen through multivariate empirical dynamical quantiles. Tables included which specify number of representative in each state or territory.

land area, the number of ORDL stay relatively stable, but overall percentage decreases. Changes the number of ORDL change the number representatives in each state/territory, but not proportionately to the size of the spatial partition. This shows evidence of the climate variability of different parts of Australia. As a country, Australia has an extremely variable climate [49, 55]. Specifically, investigated in [56], under the Köppen climate classification, northern Australia has a lot more variability in its climate. This results in more ORDL being needed to fully represent the true diversity in the nature of the system in those regions. Whereas southern Australia, which has less climate variability, requires fewer ORDL to represent the underlying dynamics of variability in the precipitation process.

5. Conclusion

We developed a method for finding a set of optimally representative dynamical locations in multivariate spatiotemporal data. This was achieved by first defining a linear multivariate ordering function through statistical depth and eigenvalue decomposition, then generalising the empirical dynamical quantiles technique to a multivariate setting.

To demonstrate the efficacy of our method, we applied it to regional scale precipitation data, here for the entire Australian continent. Data from in-situ and multi-source satellite platforms were analysed. By performing data fusion, we were able to demonstrate the influences of measurement accuracy and spatial density in the data through changes in the resulting ORDL.

We explored two applications of ORDL that can help support decision-making from data. The first lies in identifying parts of the studied region which have endured the most significant variability as opposed to those which have remained relatively stable. The second lies the future design and planning of sensor networks in different contextual domains where it is essential to account for the dynamics of several contributing factors under a changing climate. Such factors include not just the observed property (e.g., precipitation) but also that of the target event – whether this is a natural hazard to be mitigated (e.g. landslides), a human activity (e.g. agricultural crop planning and management) to be optimised, or a coupled human-natural process (e.g. soil carbon flux) [57].

Finally, although forecasting is outside the scope of the present study, our method for finding ORDL offers an important advance for forecasting from big multivariate data since we can use ORDL not only to forecast at

these critical locations but also impute missing precipitation values in data sparse areas. We have already demonstrated this process in the context of big univariate time series data from landslide monitoring [17]. This work opens the way for forecasting from multivariate spatiotemporal data. This is the subject of ongoing work.

Acknowledgement

The authors thank Yuriy Kuleshov of the Bureau of Meteorology for provision of Australian rain gauge data. We also thank the U.S. Army International Technology Center Pacific (ITC-PAC) and US DoD High Performance Computing Modernization Program (HPCMP) under Contract No. FA5209-18-C-0002 for financial support.

References

[1] M. H. Dore, Climate change and changes in global precipitation patterns: what do we know?, Environment international 31 (8) (2005) 1167–1181.

[2] J. Alcamo, T. Henrichs, T. Rosch, World water in 2025, World water series report 2 (2000).

[3] L. Rosa, D. D. Chiarelli, M. C. Rulli, J. Dell'Angelo, P. D'Odorico, Global agricultural economic water scarcity, Science Advances 6 (18) (2020) eaaz6031.

[4] M. Zeitoun, The global web of national water security, Global Policy 2 (3) (2011) 286–296.

[5] K. Bakker, Water security: research challenges and opportunities, Science 337 (6097) (2012) 914–915.

[6] M. Kotz, A. Levermann, L. Wenz, The effect of rainfall changes on economic production, Nature 601 (7892) (2022) 223–227.

[7] C. Cook, K. Bakker, Water security: Debating an emerging paradigm, Global environmental change 22 (1) (2012) 94–102.

[8] O. E. Sala, L. A. Gherardi, L. Reichmann, E. Jobbagy, D. Peters, Legacies of precipitation fluctuations on primary production: theory and data

synthesis, Philosophical Transactions of the Royal Society B: Biological Sciences 367 (1606) (2012) 3135–3144.

[9] M. Turco, J. von Hardenberg, A. AghaKouchak, M. C. Llasat, A. Provenzale, R. M. Trigo, On the key role of droughts in the dynamics of summer fires in mediterranean europe, Scientific reports 7 (1) (2017) 1–10.

[10] H. Madsen, D. Lawrence, M. Lang, M. Martinkova, T. Kjeldsen, Review of trend analysis and climate change projections of extreme precipitation and floods in europe, Journal of Hydrology 519 (2014) 3634–3650.

[11] J. A. Coe, J. A. Michael, R. A. Crovelli, W. Z. Savage, W. T. Laprade, W. D. Nashem, Probabilistic assessment of precipitation-triggered landslides using historical records of landslide occurrence, seattle, washington, Environmental & Engineering Geoscience 10 (2) (2004) 103–122.

[12] G. Myhre, K. Alterskjær, C. W. Stjern, Ø. Hodnebrog, L. Marelle, B. H. Samset, J. Sillmann, N. Schaller, E. Fischer, M. Schulz, et al., Frequency of extreme precipitation increases extensively with event rareness under global warming, Scientific reports 9 (1) (2019) 1–10.

[13] D. Kirschbaum, C. S. Watson, D. R. Rounce, D. H. Shugar, J. S. Kargel, U. K. Haritashya, P. Amatya, D. Shean, E. R. Anderson, M. Jo, The state of remote sensing capabilities of cascading hazards over high mountain asia, Frontiers in Earth science (2019) 197.

[14] V. Gupta, V. Singh, M. K. Jain, Assessment of precipitation extremes in india during the 21st century under ssp1-1.9 mitigation scenarios of cmip6 gcms, Journal of Hydrology 590 (2020) 125422.

[15] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, Y. Zhu, Big data for remote sensing: Challenges and opportunities, Proceedings of the IEEE 104 (11) (2016) 2207–2219.

[16] A. Labrinidis, H. V. Jagadish, Challenges and opportunities with big data, Proceedings of the VLDB Endowment 5 (12) (2012) 2032–2033.

[17] H. Wang, G. Qian, A. Tordesillas, Modeling big spatio-temporal geo-hazards data for forecasting by error-correction cointegration and dimension-reduction, Spatial Statistics (2020) 100432.

[18] D. Peña, R. S. Tsay, R. Zamar, Empirical dynamic quantiles for visualization of high-dimensional time series, Technometrics (2019) 1–25.

[19] F. M. Woldemeskel, B. Sivakumar, A. Sharma, Merging gauge and satellite rainfall with specification of associated uncertainty across australia, Journal of Hydrology 499 (2013) 167–176.

[20] F. J. Tapiador, F. J. Turk, W. Petersen, A. Y. Hou, E. García-Ortega, L. A. Machado, C. F. Angelis, P. Salio, C. Kidd, G. J. Huffman, et al., Global precipitation measurement: Methods, datasets and applications, Atmospheric Research 104 (2012) 70–97.

[21] V. Chakrapani Lekha, T. Oommen, S. Chatterjee, K. Sajinkumar, Rain gauge optimization for network expansion in a data-sparse region, in: First International Meeting for Applied Geoscience & Energy, Society of Exploration Geophysicists, 2021, pp. 3086–3090.

[22] A. K. Mishra, P. Coulibaly, Developments in hydrometric network design: A review, Reviews of Geophysics 47 (2) (2009).

[23] E. Pardo-Igúzquiza, Optimal selection of number and location of rainfall gauges for areal rainfall estimation using geostatistics and simulated annealing, Journal of hydrology 210 (1-4) (1998) 206–220.

[24] I. Australia, State of australian cities 2010, Infrastructure Australia Major Cities Unit (2010) 13–15.

[25] R. Koenker, G. Bassett Jr, Regression quantiles, Econometrica: journal of the Econometric Society (1978) 33–50.

[26] R. Serfling, Quantile functions for multivariate analysis: approaches and applications, Statistica Neerlandica 56 (2) (2002) 214–232.

[27] P. C. Mahalanobis, On the generalized distance in statistics, National Institute of Science of India, 1936.

[28] M. R. Neto, The concept of depth in statistics, Tech. rep. (2008).

[29] K. Mosler, Depth statistics, in: Robustness and complex data structures, Springer, 2013, pp. 17–34.

[30] Y. Zuo, R. Serfling, General notions of statistical depth function, Annals of statistics (2000) 461–482.

[31] H. Dette, S. Hoderlein, N. Neumeyer, Testing multivariate economic restrictions using quantiles: the example of slutsky negative semidefiniteness, Journal of Econometrics 191 (1) (2016) 129–144.

[32] R. Serfling, A depth function and a scale curve based on spatial quantiles, in: Statistical data analysis based on the L1-norm and related methods, Springer, 2002, pp. 25–38.

[33] J. D. Hamilton, Time series analysis, Vol. 2, Princeton university press Princeton, NJ, 1994.

[34] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometrics and intelligent laboratory systems 2 (1-3) (1987) 37–52.

[35] A. Ukkola, M. Roderick, A. Barker, A. Pitman, Exploring the stationarity of australian temperature, precipitation and pan evaporation records over the last century, Environmental Research Letters 14 (12) (2019) 124035.

[36] J. Bloomfield, B. Marchant, Analysis of groundwater drought building on the standardised precipitation index approach, Hydrology and Earth System Sciences 17 (12) (2013) 4769–4787.

[37] D. Tigkas, H. Vangelis, G. Tsakiris, Drought characterisation based on an agriculture-oriented standardised precipitation index, Theoretical and applied climatology 135 (3) (2019) 1435–1447.

[38] C. K. Wikle, A. Zammit-Mangion, N. Cressie, Spatio-temporal Statistics with R, Chapman and Hall/CRC, 2019.

[39] L. Chao, K. Zhang, Z. Li, Y. Zhu, J. Wang, Z. Yu, Geographically weighted regression based methods for merging satellite and gauge precipitation, Journal of Hydrology 558 (2018) 275–289.

[40] H. Wu, Q. Yang, J. Liu, G. Wang, A spatiotemporal deep fusion model for merging satellite and gauge precipitation in china, Journal of Hydrology 584 (2020) 124664.

[41] L. Xu, N. Chen, H. Moradkhani, X. Zhang, C. Hu, Improving global monthly and daily precipitation estimation by fusing gauge observations, remote sensing, and reanalysis data sets, Water Resources Research 56 (3) (2020) e2019WR026444.

[42] S. Ochoa-Rodriguez, L.-P. Wang, P. Willems, C. Onof, A review of radar-rain gauge data merging methods and their potential for urban hydrological applications, Water Resources Research 55 (8) (2019) 6356–6391.

[43] B. Hines, Y. Kuleshov, G. Qian, Spatial modelling of linear regression coefficients for gauge measurements against satellite estimates, 2019-20 MATRIX Annals (2018).

[44] Q. Fu, R. Ruan, Y. Liu, Accuracy assessment of global satellite mapping of precipitation (gsmap) product over poyang lake basin, china, Procedia Environmental Sciences 10 (2011) 2265–2271.

[45] Z.-W. Chua, Y. Kuleshov, A. Watkins, Evaluation of satellite precipitation estimates over australia, Remote Sensing 12 (4) (2020) 678.

[46] R. Pearson, Limnology in the northeastern tropics of australia, the wettest part of the driest continent, Internationale Vereinigung für Theoretische und Angewandte Limnologie: Mitteilungen 24 (1) (1994) 155–163.

[47] S. Jiang, L. Ren, Y. Hong, X. Yang, M. Ma, Y. Zhang, F. Yuan, Improvement of multi-satellite real-time precipitation products for ensemble streamflow simulation in a middle latitude basin in south china, Water resources management 28 (8) (2014) 2259–2278.

[48] G. Tang, M. P. Clark, S. M. Papalexiou, Z. Ma, Y. Hong, Have satellite precipitation products improved over last two decades? a comprehensive comparison of gpm imerg with nine satellite and reanalysis datasets, Remote sensing of environment 240 (2020) 111697.

[49] N. Nicholls, W. Drosdowsky, B. Lavery, Australian rainfall variability and change, Weather 52 (3) (1997) 66–72.

[50] I. Smith, An assessment of recent trends in australian rainfall, Australian Meteorological Magazine 53 (3) (2004) 163–173.

[51] S. K. Adhikary, A. G. Yilmaz, N. Muttil, Optimal design of rain gauge network in the middle yarra river catchment, australia, Hydrological processes 29 (11) (2015) 2582–2599.

[52] E. Bernard, P. Naveau, M. Vrac, O. Mestre, Clustering of maxima: Spatial dependencies among heavy rainfall in france, Journal of climate 26 (20) (2013) 7929–7937.

[53] S. Doocy, A. Daniels, S. Murray, T. D. Kirsch, The human impact of floods: a historical review of events 1980-2009 and systematic literature review, PLoS currents 5 (2013).

[54] M. N. Halgamuge, A. Nirmalathas, Analysis of large flood events: Based on flood data during 1985–2016 in australia and india, International journal of disaster risk reduction 24 (2017) 1–11.

[55] J. S. Risbey, M. J. Pook, P. C. McIntosh, M. C. Wheeler, H. H. Hendon, On the remote drivers of rainfall variability in australia, Monthly Weather Review 137 (10) (2009) 3233–3253.

[56] L. Tang, F. Hossain, Investigating the similarity of satellite rainfall error metrics as a function of köppen climate classification, Atmospheric Research 104 (2012) 182–192.

[57] P. J. Ferraro, J. N. Sanchirico, M. D. Smith, Causal inference in coupled human and natural systems, Proceedings of the National Academy of Sciences 116 (12) (2019) 5311–5318.