Sustainable Cities and Society 27 (2016) 484–496

Contents lists available at ScienceDirect

Sustainable Cities and Society

journal homepage: www.elsevier.com/locate/scs

Cluster analysis of urban water supply and demand: Toward large-scale comparative sustainability planning

Karen Noiva*, John E. Fernández, James L. Wescoat Jr.

Department of Architecture, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 29 January 2016
Received in revised form 2 June 2016
Accepted 4 June 2016
Available online 7 June 2016

Keywords:
Cluster analysis
Comparative methods
Sustainable urban water systems
Water budget
Water use

ABSTRACT

The sustainability of urban water systems is often compared in small numbers of cases selected as much for their familiarity as for their similarities and differences. Few studies examine large urban datasets to conduct comparisons that identify unexpected similarities and differences among urban water systems and problems. This research analyzed a dataset of 142 cities that includes annual per capita water use (m^3/yr/cap) and population. It added a 0.5° grid annual water budget value (P-PET/yr) as an index of hydroclimatic water supply. With these indices of urban water supply and demand, we conducted a hierarchical cluster analysis to identify relative similarities among, and distances between, the 142 cases. While some expected groupings of climatically similar cities were identified, unexpected clusters were also identified, e.g., cities that use water at greater rates than local climatic water budgets provide. Those cities must seek water from greater distances and greater depths. They face greater water and wastewater treatment costs. To become more sustainable they must increase water use efficiency, demand management, reuse, and recycling. The significance of the population variable suggests that adding other explanatory socio-economic variables, as well as more precise water system indices, are logical next steps for comparative analysis of urban water sustainability.

1. Introduction and conceptual framework

Comparative urban water research is important for drawing generalizations about sustainability and for adapting lessons from one set of cities for consideration in others. The current status of comparative urban water research is limited (Mollinga & Gondhalekar, 2014; Wescoat, 2014). Many studies examine single case studies (e.g., Gandy, 2014). Some select cases that reflect the authors' expertise (Novotny, Ahern, & Brown, 2010; Wörlen et al., 2016). Others group case studies under predefined headings, e.g., arid, tropical, low-income, megacities, etc. (Fletcher & Deletic, 2008). Most strive for generalizable models and methods but without analyzing a large number of cases (Mollinga & Gondhalekar, 2014). As a result, comparisons and the conclusions that can be drawn from them tend to be limited and qualitative (Mollinga & Gondhalekar, 2014; Wescoat, 2009; Wescoat, 2014). This study lays a foundation for comparative research on urban water sustainability using cluster analysis methods.

The sustainability framework employed here involves two simple mass balance variables for urban water systems. The first is an annual climatic water budget for each urban area, which subtracts potential evaporation from precipitation (elaborated in the methods section below) (Willmott, Rowe, & Mintz, 1985). Water balance analysis estimates climatic water supply (P) and demand (PET). Some cities have an annual water balance surplus that can be stored or managed as runoff, while others have climatic water deficits that must be managed through rainwater harvesting, water use efficiency, recycling, reuse and, barring those methods, long distance water imports (Plappally & Lienhard, 2012; Plappally & Lienhard, 2013). The second metric is gross annual water use per capita in each city. These two variables provide annual estimates of climatic water supply and gross per capita water demand. While one would expect some correlation between supply and demand, cities have historically supplemented local water supplies with long distance water transfers, aquifer depletion, and in some cases desalination (McDonald, Weber, Padowski, Flörke, & Schneider, 2014). These variables are weakly correlated and are treated here as independent variables to classify the sustainability of urban water patterns.

The water balance approach may be compared with other sustainability heuristics such as water footprint analysis, which examines different types and amounts of water use in a system (Hoekstra and Chapagain, 2007; Hoff, Döll, Fader, Gerten, & Hauser, 2014). Here we adapt the footprint idea in an urban Water Use and Climate Index (WUCI in m^2/cap) (See Section 2). This type

* Corresponding author.
E-mail addresses: knoiva@mit.edu, knoiva@gmail.com (K. Noiva).

http://dx.doi.org/10.1016/j.scs.2016.06.003

of accounting is taken further in studies of virtual water trade and sustainable supply chain analysis (e.g., Daniels, Lenzen, & Kenway, 2011; Ercin, Aldaya, & Hoekstra, 2011; Hoff et al., 2014; Konar, Dalin, Hanasaki, Rinaldo, & Rodriguez-Iturbe, 2012; Suweis, Konar, Dalin, Hanasaki, & Rinaldo, 2011). Explanatory sustainability heuristics such as the IPAT equation (I = P*A*T) relate environmental impacts (e.g. resource use) (I) to population size (*P*), affluence (*A*), and technology (T) (Rosa, York, & Dietz, 2004). The IPAT equation takes multiple forms. One popular version is $I = P^*F$, where *F* is impact per capita (Chertow, 2001). While IPAT is presented as an equation, it is more of a heuristic of driving and ameliorating factors (Chertow, 2001). Greater emphasis on explanatory analysis of water supply and demand patterns is needed in each approach, but a first step toward that aim is identifying and characterizing urban water patterns, which can then enable systematic subsampling and comparison.

Application of statistical data-mining techniques such as cluster analysis to urban socio-economic classification is well-established (Astel, Tsakovski, Barbieri, & Simeonov, 2007; Bettencourt, 2013; Kim, 1997; Batty, Axhausen, Giannotti, Pozdnoukhov, & Bazzani, 2012; MacCannell, 1957; Kennedy, 2011). Application to water issues within cities is a more recent development (Yu et al., 2013; Diao, Farmani, Fu, Astaraie-Imani, & Ward, 2014), as is classification of urban water systems within a particular country (Yu & Chen, 2010; Rahill-Marier & Lall, 2013; Rao & Srinivas, 2008a, Chapter 3; Rao & Srinivas, 2008b, Chapter 3). In reviewing the water resources literature, Mollinga and Gondhalekar suggested an approach for assessing 'small-N' and 'medium-N' case studies (Mollinga & Gondhalekar, 2014). For example, clustering has been applied to the problem of forecasting short-term water demand within a single city or municipal water system, which falls under the category of a small-N analysis (Garg, 2007; Candelieri & Archetti, 2014; Wu, Lv, Dong, Wang, & Xu, 2012). Other clustering studies fall into the medium-N category of comparative analysis, including one that includes a *k*-means clustering of cities based on water footprint, energy consumption, and municipal waste within the United Kingdom (Khamis, 2012). A study by Mayer et al. used cluster analysis to classify watersheds in the Great Lakes basin according to social and environmental attributes (Mayer, Winkler, & Fry, 2014). And a large-N study by the Columbia University Water Center used hierarchical clustering to analyze utility rates in the United States with regards to financial sustainability (Rahill-Marier & Lall, 2013).

This paper uses clustering algorithms to analyze water supply and demand for 142 cities around the world to develop an international classification of urban water sustainability situations. It uses the MIT Urban Metabolism database (Fig. 1), which was created to develop an urban sustainability typology based on four predictor variables (population, population density, affluence [GDP per capita], and climate [Köppen classification]); and eight response variables (per capita consumption of construction minerals, industrial minerals, biomass, water, total energy, total materials, fossil fuels, and electricity) (Saldivar-Sali, 2010; Ferrão & Fernández, 2013, Chapter 4). Our analysis uses the same set of cities and expands the large-N analysis of urban water issues as elaborated in the next section.

2. Data and methods

The Urban Metabolism dataset includes 142 cities distributed across major continents and climates. We focused parsimoniously on two variables in the UrbMet database, added a water budget variable, and then used them to construct a Water Use and Climate Index (WUCI). These variables were:

1. Per capita water consumption (CONS), to assess the scale of urban water use.
2. Population (POP), to assess the potential significance of city size.
3. Net annual climatic water budget (DIFF), to assess hydroclimatic water supplies.
4. Water Use and Climate Index (WUCI), to provide a measure of urban water use per capita indexed to annual precipitation.

The main water variable in the UrbMet dataset is per capita water use in cubic meters per year (m^3/cap/year) (Table 1). This value was drawn in most cases from the World Bank-supported International Benchmarking Network (IBNET) supplemented by city-specific data when not available in IBNET (Saldivar-Sali, 2010; IBNET, 2015). The definition of per capita water use in IBNET is gross annual water production by a utility divided by the number of people in the service area. Per capita water use ranged from 14 m^3/cap/year (Yangon) to 355 m^3/cap/year (Cairo). We did not disaggregate per capita use into residential and commercial subsectors, though that would be a valuable extension of this research.

Climatic water balance analysis was employed to estimate gross annual water supplies, using the University of Delaware's 0.5° grid WebWIMP[1] tool (Willmott et al., 1985). The annual Difference (DIFF) between Precipitation and Potential Evapotranspiration, in meters per year (m/yr), for each of the 142 cities was scraped from the database. The 0.5° resolution was coarse, but deemed appropriate for indexing the water balance of major urban areas, which generally occupy and draw upon larger catchment areas outside their administrative boundaries. Cities in the database had DIFF values ranging from −1.446 m/yr (Abu Dhabi) to +3.833 m/yr (Anchorage).

City population data (POP) from the UrbMet database was included to assess the potential difference that city size makes for classifying patterns of urban water supply and demand. Population size ranges from 270,000 (Bandar Seri Begawan) to 14.34 million (Shanghai). These population estimates from the early 2000s are conservatively based on city boundaries vis-à-vis larger metropolitan regions.

The aforementioned WebWIMP tool also provided average annual precipitation data (PREC) in meters per year for each city. PREC[2] was used to calculate Water Use and Climate Index (WUCI), which had units of m^2/capita:

$$\mathrm{WUCI} = \mathrm{CONS}/\mathrm{PREC}$$

WUCI was not included as a metric in the cluster analysis, but was used in visualizing the results.

Several methods were used to visualize individual urban water variables and relationships among them. Prior to clustering, we explored the data using qq-normal plots, histograms, and scatterplots. We first plotted data for each metric on a bar chart where the cities are sorted according to their value for that metric (Fig. 2a–c). Fig. 2a and b have a similar distribution of positive values, with a longer tail of smaller cities in Fig. 2a and a more even progression of water consumption values in Fig. 2b. In contrast, Fig. 2c includes negative as well as escalating positive values for net annual water balance (Fig. 3).

Prior to clustering, these variables were transformed using a base-10 logarithmic transformation (log10), which reduced skewness and improved the symmetry of their distribution (Fig. 4a and

[1] WebWIMP is short for the "Web-based, Water-Budget, Interactive, Modeling Program" and is available through the University of Delaware at: http://climate.geog.udel.edu/~wimp/.

[2] PREC was used instead of DIFF in calculating WUCI, since DIFF had values close to zero and therefore could not be used in the denominator. However, PREC was highly correlated to DIFF.

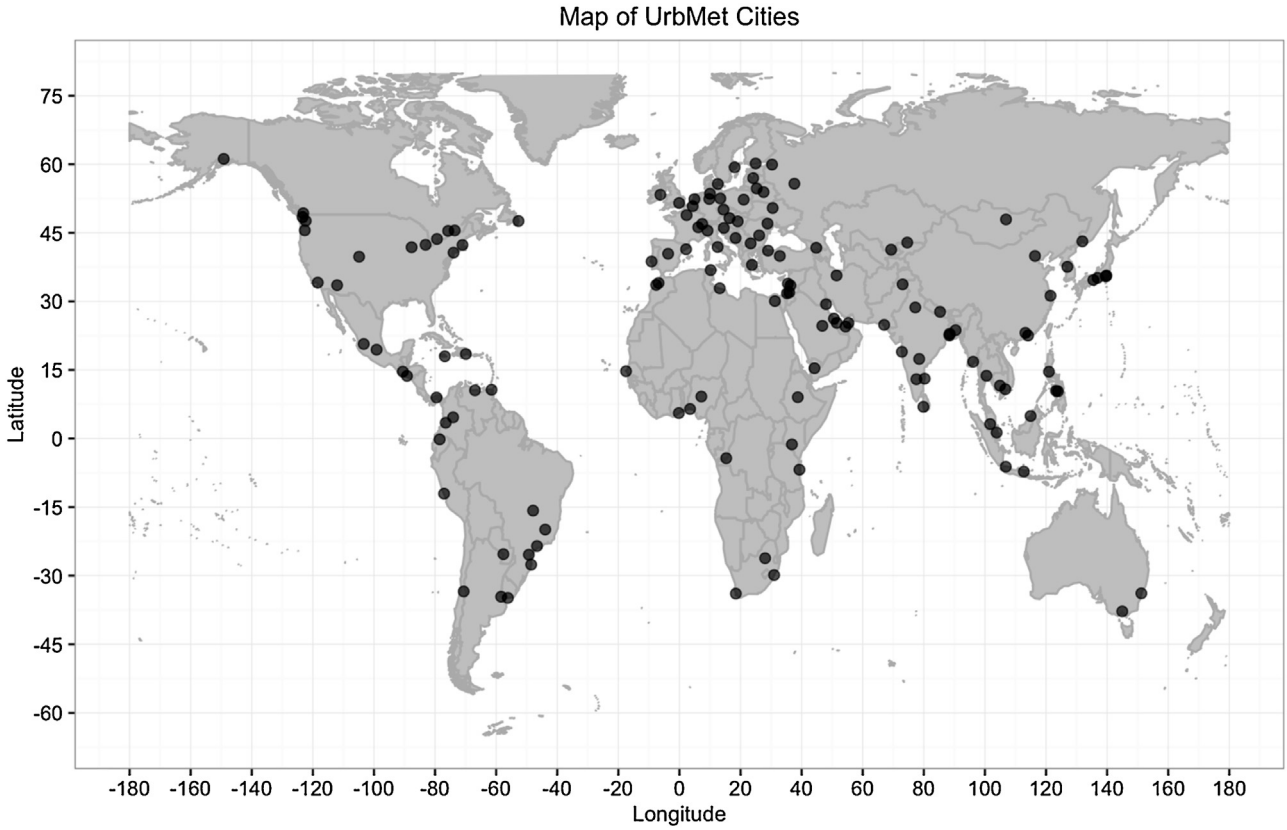

Fig. 1. World map of the 142 cities in the UrbMet database.

Table 1
Descriptive statistics for CONS, DIFF, POP, and WUCI for the UrbMet database.

Metric	Unit	Quartile Break						
		0%	25%	50%	75%	100%	Mean	St. Dev.
CONS	m^3/cap/year	14.0	57.3	86.5	148.5	355.0	110.8	75.2
DIFF	m./year	−1.446	−0.241	0.089	0.447	3.833	0.124	0.752
POP	millions	0.0273	0.647	1.649	3.764	14.350	2.883	3.182
WUCI	m^2/cap	5.717	49.04	117.9	189.1	14200.0	303.7	1223.0

b). A constant, a, was added to DIFF prior to the transform such that $a = 1 - \text{DIFF}_{\text{min}}$ (where DIFF_{min} was the minimum value of DIFF).

While it might be expected to find correlation between the size of the population or water consumption and local water availability (i.e., DIFF), the correlation was found to be low. As seen in Table 2, the r value for log10(CONS) vs. log10(DIFF + a) was found to be 0.02 with a significance level of $p = 0.8373$, while the r value for log10(POP) vs. log10(DIFF + a) was found to be 0.01 with a $p = 0.9261$. In other words, the log10 transforms of CONS and DIFF were found to be independent, as were the log10 transforms of DIFF and POP. A small negative correlation of $r = -0.12$ was found between log10(CONS) vs. log10(POP), but the significance level was only 0.1684. In other words, these three variables seem to be independent of each other, justifying an exploratory data-mining approach to the data.

2.1. Cluster analysis methods

Cluster analysis is used in exploratory data mining to group objects in a dataset in such a way that those within a group are more similar to each other than to objects in other groups. Common clustering approaches include hierarchical clustering, k-means clustering, and model-based clustering. For purposes of exploring these data, hierarchical clustering was chosen. We used t-Distributed Stochastic Neighbor Embedding to reduce dimensionality in the data for visualization and clustering (van der Maaten & Hinton, 2008). After the log10 transformation each metric was scaled to unity. A distance matrix was then calculated using the **dist** function from the **stats** library in R and the Euclidean distance formula in which $||a - b||_2 = \text{SQRT}(\text{SUM}(a_i - b_i)^2)$. The basic Euclidean distance formula was used as there were no theoretical reasons to prefer a more complex formula, and other formulas did not produce substantially different or more interesting results. These visualization methods are briefly described below and displayed as Figs. 6–8 in the results section.

2.1.1. t-Distributed Stochastic Neighbor Embedding in scatterplots

The t-SNE method for reducing dimensionality enhances visualization in scatterplots by iteratively assigning each high-dimensional object to a point in a two-dimensional space. The points are assigned such that neighbors are more similar to each other than to distant objects. In a two-dimensional variable space, the human eye can to some extent distinguish clustered groups of observations from one another. As dimensionality increases, the task becomes substantially more difficult and less intuitive. In

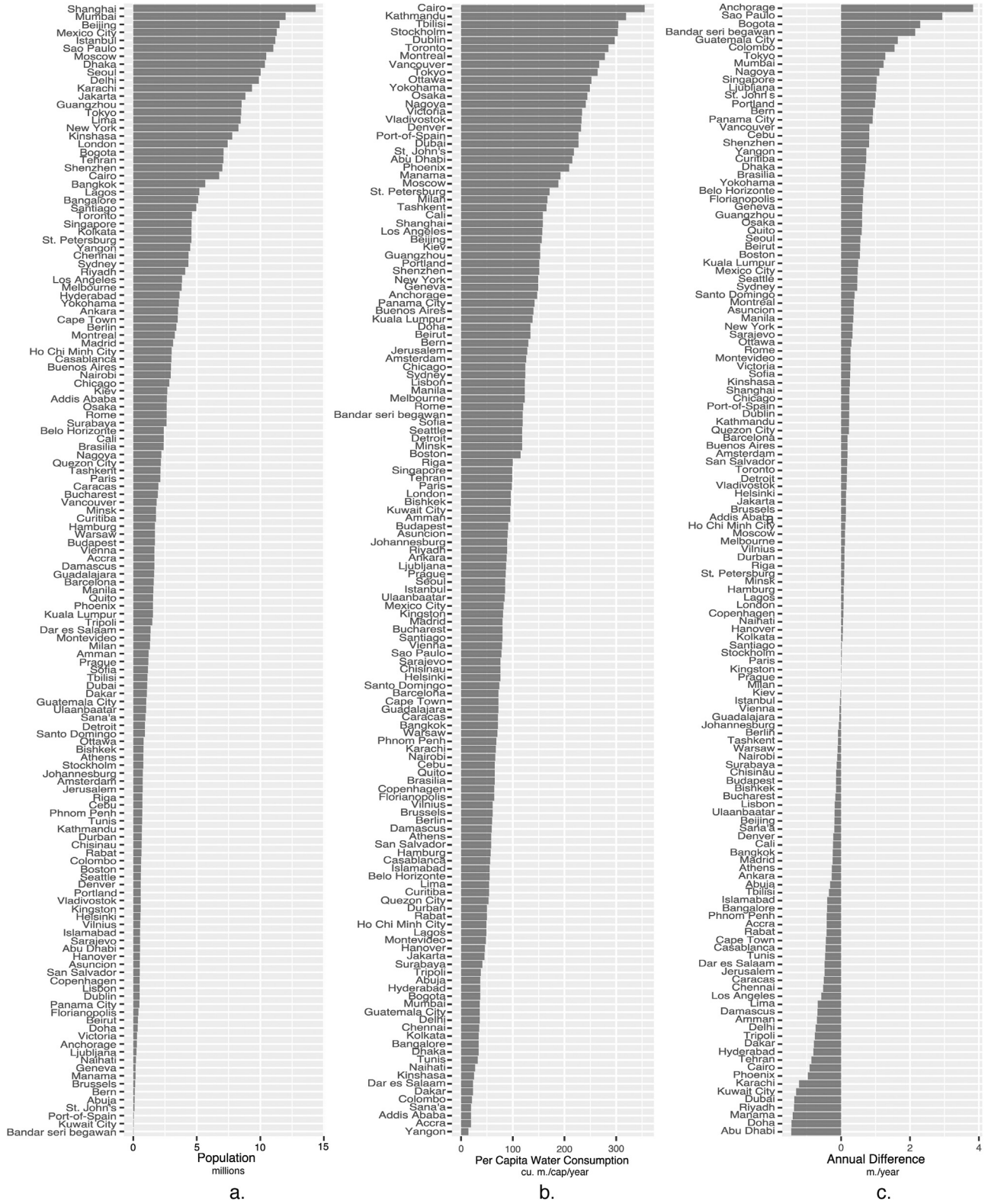

Fig. 2. Population (millions), per capita water consumption (cu. m./cap/year), and annual difference (calculated as the difference (DIFF) between precipitation and potential evapotranspiration in m/year).

reducing dimensionality, the t-SNE approach enables observations to be visualized in a way that is more intuitive to the human eye.

The distance matrix produced by **dist** was used as the dissimilarity matrix given as the argument to t-SNE*(t-Distributed Stochastic Neighbor Embedding).*t-SNE produced two vectors containing the coordinates embedding each city within the t-SNE space. These vectors were then scaled to unity and a second distance matrix was constructed from these data. This second distance matrix was then passed as an argument to the hierarchical clustering algorithm **hclust,** which takes an agglomerative approach to hierarchical clus-

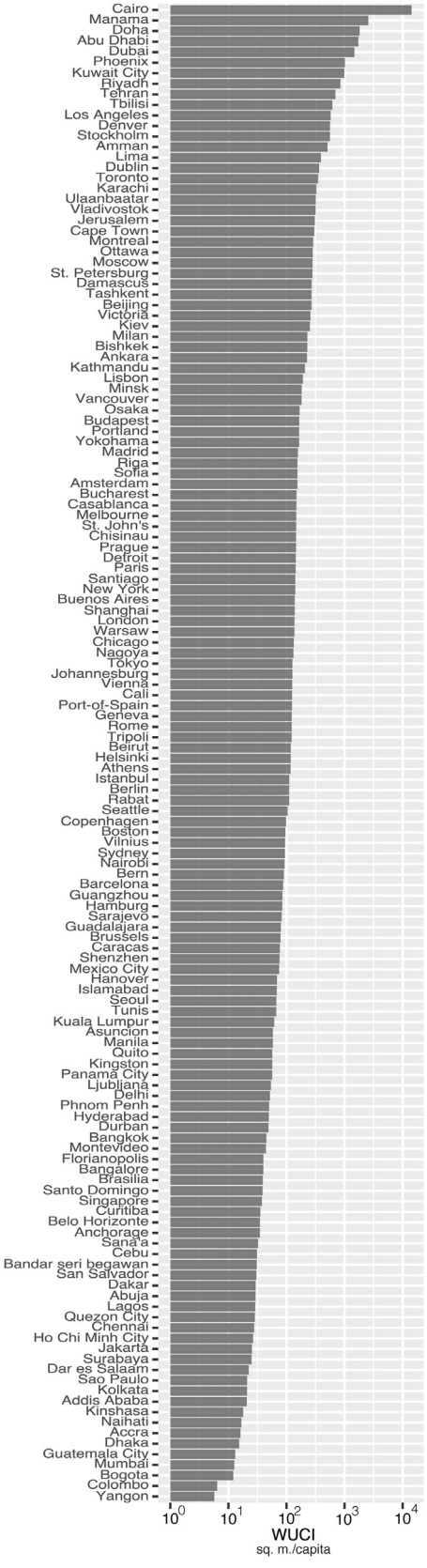

Fig. 3. Water use and climate index in m^2/capita (WUCI = CONS/PREC).

Table 2
Pearson's correlation coefficient and significance of base-10 logarithmic transformations of average annual per capita water consumption (CONS), average annual difference (DIFF), and population (POP).

Metric	log10(CONS)	log10(DIFF + a)	log10(POP)
log10(CONS)	1.00	0.02, $p = 0.8373$	−0.12, $p = 0.1684$
log10(DIFF + a)		1.00	0.01, $p = 0.9261$
log10(POP)			1.00

tering. Each observation starts in its own cluster, and pairs of clusters are joined at each iterative step. At each iterative step, the **hclust** groups are compared based on a linkage criterion. Several common linkage criteria exist. We used Ward's minimum variance method, which minimizes the squared Euclidean distance. This criterion led to a decrease in variance for the cluster being merged: at each step, the pair of clusters merged is based on the optimal value of the error sum of squares ($d_{ij} = d(\{X_i\}, \{X_j\}) = || X_i - X_j ||^2$).

2.1.2. Dendrograms

The results of hierarchical clustering were also visualized using a tree-like structure known as a dendrogram, which represents relationships of similarity amongst the observations. The dendrogram of cluster results shows the step-wise pairing of existing subclusters. Each branch is called a "clade" and each terminal node is a "leaf". The horizontal distance of each branch indicates a measure of the distance between clades. The dendrogram is "cut" to a desired height or number of clusters to determine the membership of each leaf (i.e., city).

2.1.3. Violin plots/boxplots

The cluster results were also plotted as violin plots with boxplots superimposed over them. The violin plot is essentially a box plot with a rotated kernel density plotted instead of a box. The boxplot (or box-and-whisker plot) depicts the median value of each cluster as a band within the box. The top and bottom of each box (the "hinges") represent the first and third quartiles. Boxplots may include "whiskers"—vertical lines extending from the tops and bottoms of the boxes. The length of the whiskers is determined by the inter-quartile range (IQR), where IQR = Q3 − Q1: i.e., the difference between the third and first quartiles. The whiskers extend from each hinge to the value that is within 1.5*IQR of the hinge. Any value beyond the whiskers is an outlier and is plotted as a point.

3. Cluster analysis results

The cluster analysis yielded interesting results at several levels of division and grouping. We examined the results from two to eight clusters, and found that six clusters yielded the most varied yet still legible differences with each of the four visualization methods. The typology generated by the six clusters is presented and discussed here. The world map in Fig. 5 displays how these types appear spatially. It has some clear patterns, such as East Asian, South Asian, Central African cities, and other mixed/gradient patterns in Europe and North America. However, it has too much complexity to characterize from visual inspection alone.

The dendrogram in Fig. 6 depicts the positions of the 142 cities across the full range of clusters (i.e., from 1 to 142); cities are colored by the selected six cluster-typology. The scatterplot in Fig. 7 depicts the cities in the two-dimensional space produced using t-SNE, which conveys more clearly the relative similarity between cities and clusters.

However, the six clusters are most readily interpreted using the violin box plots shown in Fig. 8. These graphic visualizations made it possible to discern the six urban clusters in qualitative and quantitative terms. Table 3 used these figures to define the general

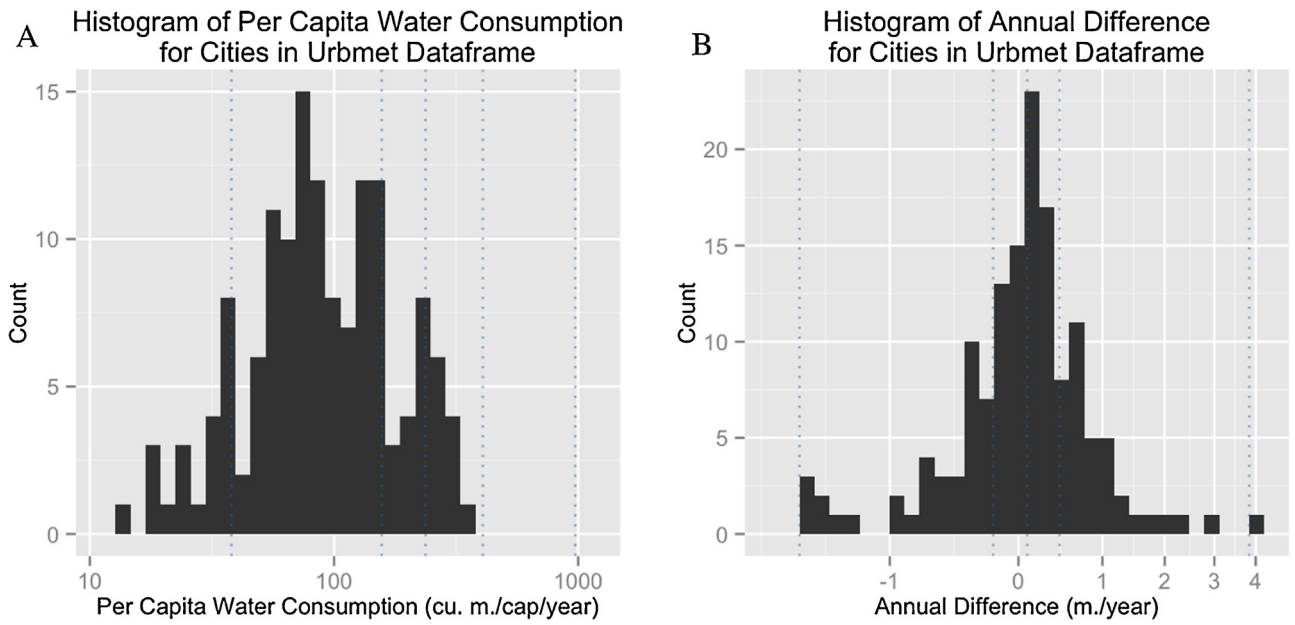

Fig. 4. Histograms for average annual per capita water consumption (CONS) (Fig. 4a); and average annual difference (DIFF) + *a* (a constant) transformed by a base-10 logarithm (Fig. 4b).

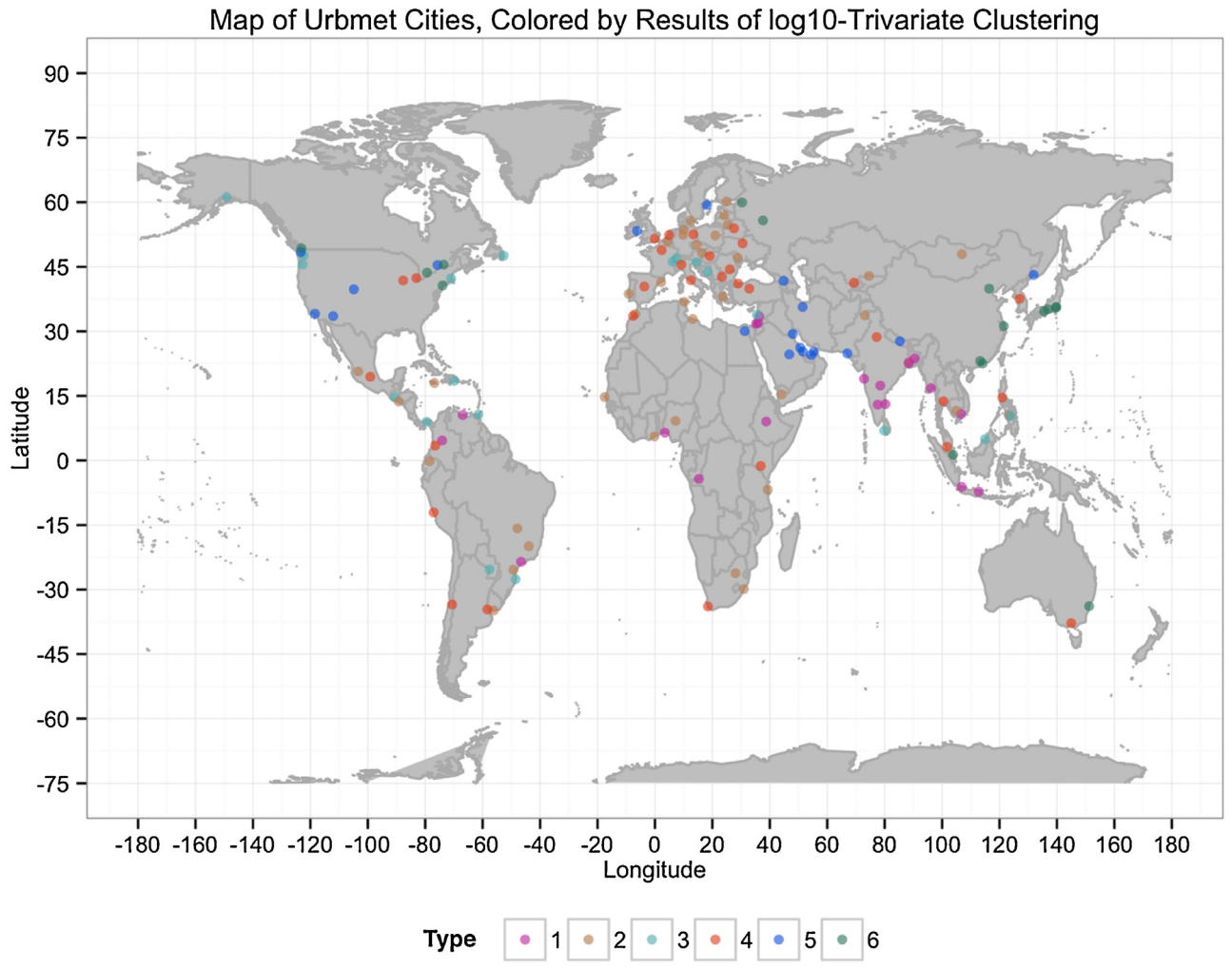

Fig. 5. The 142 cities clustered in the study, plotted on a world map and colored by type.

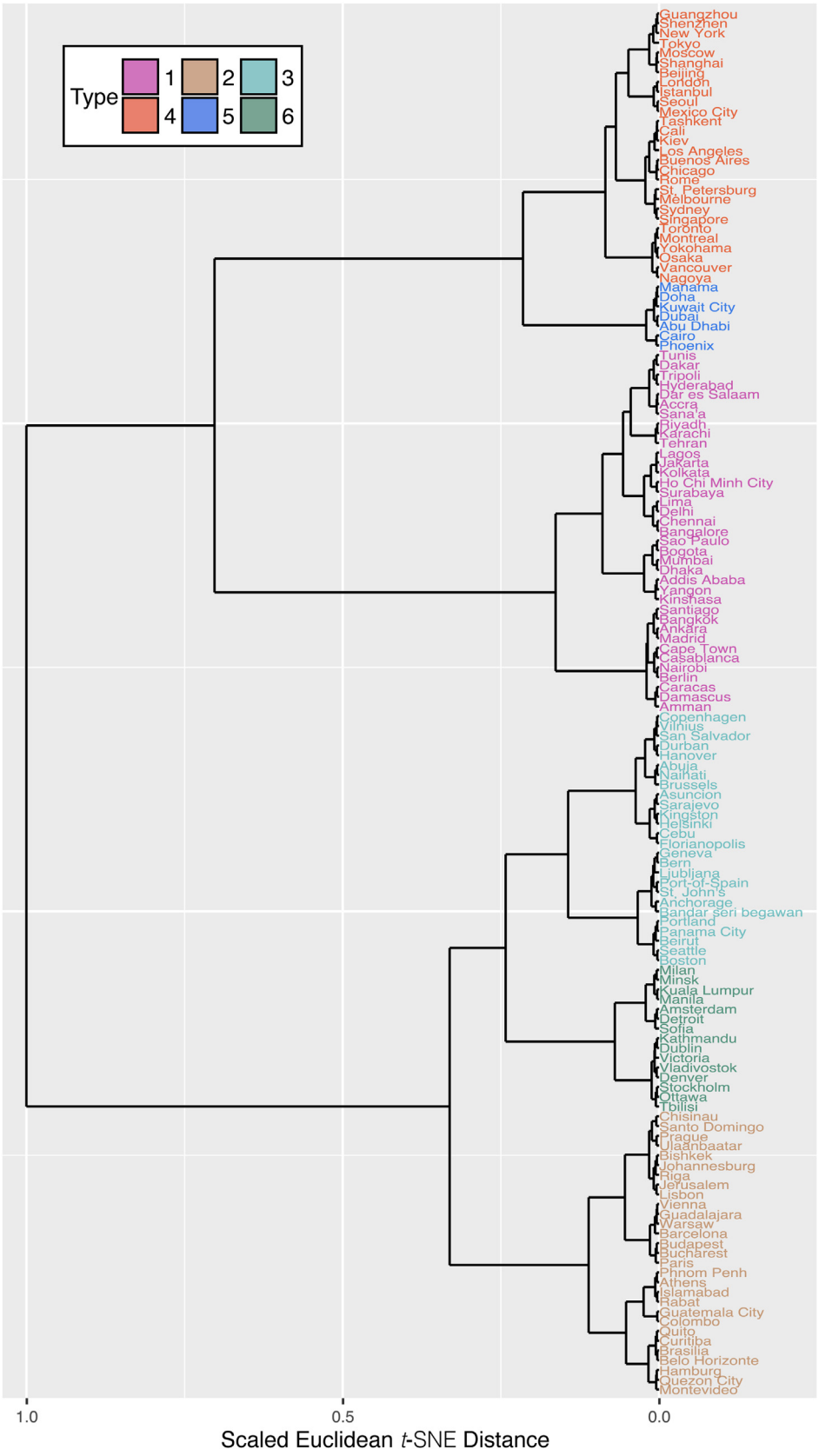

Fig. 6. Dendrogram of cluster results, with cities colored by cluster membership.

characteristics of each cluster and list several representative cities. We now recast these results as a working typology of urban water sustainability situations. The description of each "type" uses the terms "low", "medium", and "high" in quantitative as well as qualitative terms, as they are based on the quartile breaks and thresholds between clusters.

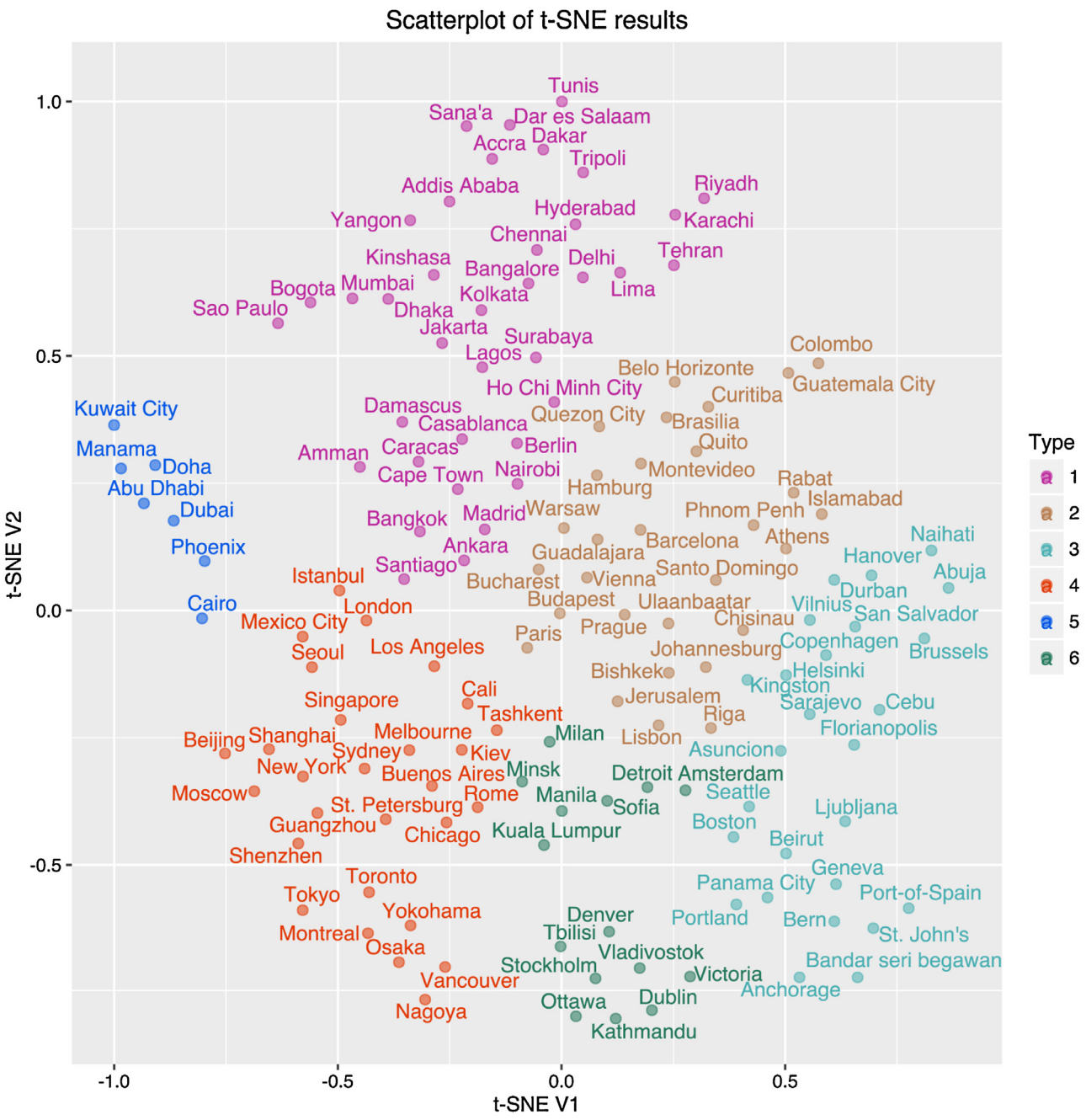

Fig. 7. Scatterplot of cities, plotted in the two-dimensional space produced by t-SNE.

3.1. Type 1: large cities with very low per capita water consumption

These cities represent a range of climate and supply conditions, but their net water balances are predominantly negative. The median value of POP for Type 1 cities lies in the 3rd quartile. Type 1 cities fall predominantly below the median value for CONS, with a median value of 45.0 m^3/cap/year, which is within the first quartile.[3] The median value for DIFF for Type 1 also lies within the 1st quartile, and the box for Type 1 on DIFF lies within the 1st and 2nd quartiles. In other words, Type 1 cities tend to have negative values for DIFF. To summarize, Type 1 cities tend to have large populations, low natural water availability, and very low per capita water consumption, which raises concerns about their sustainability in hydroclimatic and socio-economic terms. Type 1 cities are most dissimilar from cities of Type 3 and 6, which are smaller and wetter. They show some overlap with Type 2 and Type 4 cities.

3.2. Type 2: medium-sized cities with medium-low water consumption and a net water balance of around zero

Type 2 cities have values for POP that fall predominantly in the 2nd quartile, with a median value of 1.188 million; Type 2 cities tend to be medium-low to medium-sized. These cities have the second-lowest median value for CONS, after those of Type 1. Type 2 cities also have low to moderate water balances that (DIFF). In summary, Type 2 cities tend to be mid-size cities with a net water

[3] The WHO recommends around 100 L/cap/day per person (36.5 m^3/cap/year).

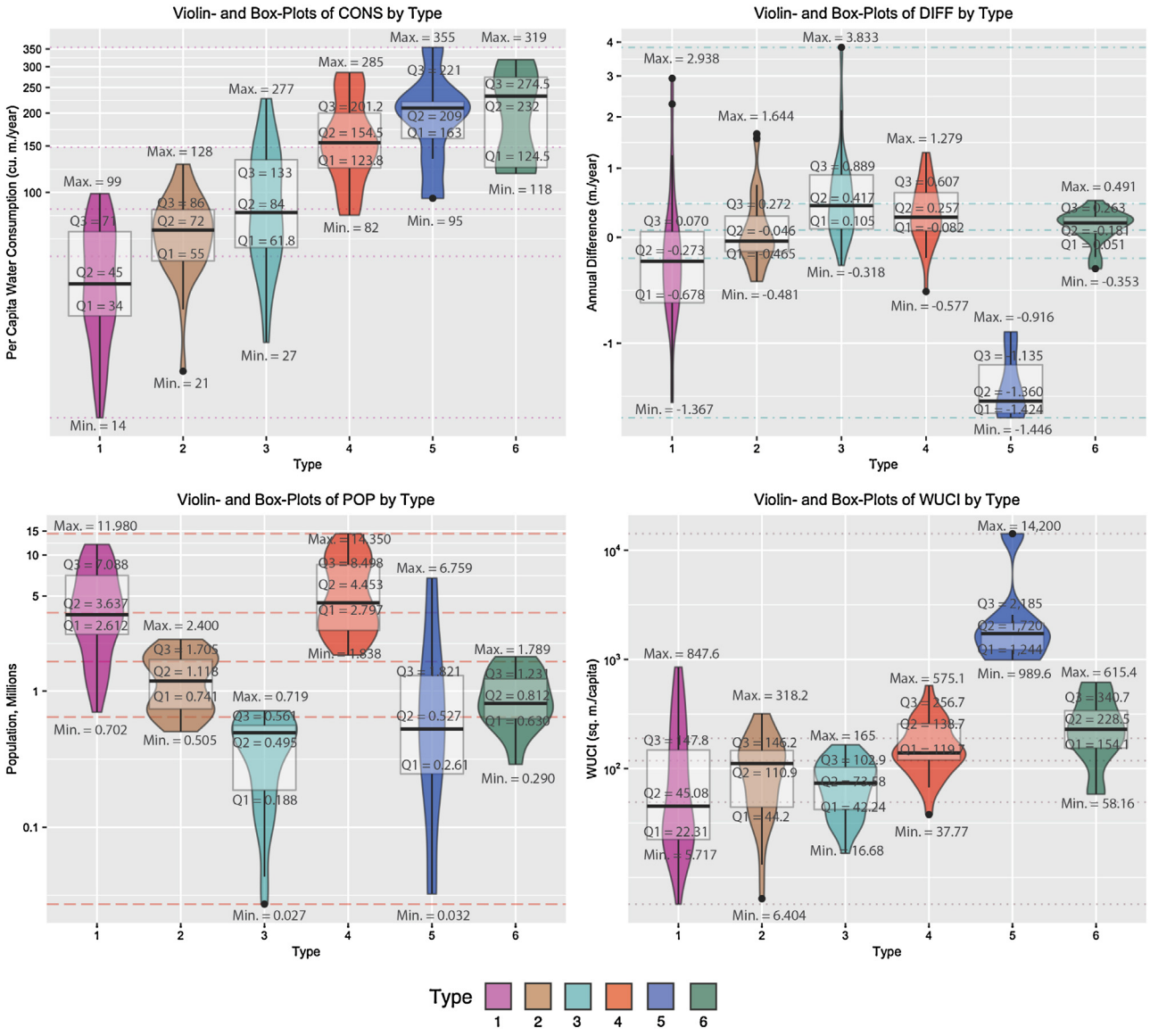

Fig. 8. Violin- and box-plots for POP, CONS, DIFF, and WUCI.

Table 3
Descriptive attributes of each urban water system Type. The designations 'high', 'medium', and 'low' are made with reference to the statistical quartiles (shown on the violin plots in dashed lines) and to the thresholds between types.

Type	DIFF	CONS	POP	WUCI	Representative Cities
1	Range from low to high but predominantly very low to low (negative).	Very low	Range from medium-low to very large	Range from low to medium, but predominantly low	Casablanca, Jakarta, Kinshasa, Lagos, Mumbai
2	Medium-low/Medium	Medium-low	Medium-sized	Range from low to medium, but predominantly low	Barcelona, Hamburg, Prague, Ulaanbaatar, Vienna
3	Range from medium to high (positive)	Medium	Generally smaller	Range from low to medium, but predominantly low	Bern, Boston, Copenhagen, Geneva, Hanover
4	Range from low to high; predominantly medium and positive	Range from medium-high to high	Range from medium-high to high; predominantly high	Range from low to medium, but predominantly medium	London, Montreal, New York, Singapore, Sydney
5	Very low	Very high	Range from very small to very large	Range from low to medium, but predominantly medium	Cairo, Doha, Dubai, Kuwait City, Phoenix
6	Medium; generally positive	Very high	Low to medium-low	High	Amsterdam, Denver, Kuala Lumpur, Milan, Stockholm

balance close to zero (i.e., moderate water resources) and medium-low water use, which are characteristics that may be sustainable.

3.3. Type 3: small cities with medium levels of water consumption and positive net water balances

These cities tend to be small relative to the other cities in the dataset; their median value is the lowest of any of the six city types at 494,700. Nearly all Type 3 cities fall within the 1 st quartile for POP. They have a median value for CONS of 84 m^3/cap/day. Type 3 cities have the highest median value for DIFF of any of the other types, at 0.417 m/year, and most Type 3 cities have values for DIFF in the 3rd and 4th quartiles. In other words, Type 3 cities tend to be small, with relatively high naturally available water resources and medium levels of water consumption. These cities have the lowest Water Use and Conservation Index (WUCI), and might therefore be deemed the most sustainable in terms of use (though not in terms of risk which is not considered in this analysis). Type 3 cities show similarities with Type 2 and Type 6; and they are most dissimilar from Types 1, 4, and 5.

3.4. Type 4: very large cities with high per capita water consumption and a positive net water balance

Type 4 cities have large populations with a median value of 4.45 million. Type 4 cities also have high per capita water use (CONS, with a median value of 154.5 m^3/cap/year, which falls in the 4th quartile). Almost all Type 4 cities are in the 3rd and 4th quartiles for CONS and DIFF. They are very large cities with large natural water supplies and high demand, and they thus have the potential to be sustainable.

3.5. Type 5: cities of varied size with very high per capita water consumption located in highly arid environments

These cities have the second-highest median value for CONS, at 209.0 m^3/cap/year; almost all of these cities have CONS that falls within the 4th quartile. However, Type 5 has the lowest DIFF of any of the types, with a median value of −1.36 m/year and a maximum value of −0.916 m/year. The POP of Type 5 cities ranges from the smallest to the largest: from 32,400 to 6.759 million. To summarize, Type 5 cities have a wide range of population size, and they are characterized by high water use and very low water budgets. This pattern raises serious concerns about their sustainability.

3.6. Type 6: medium-sized cities with very high per capita water consumption and low positive water balance

These cities tend to be medium-sized cities, with high per capita water consumption and low, positive water balances. They are one of the most rapidly growing forms of urbanization in developing countries, and this pattern raises concerns about their sustainability. Type 6 cities are most similar to Type 1, Type 2, and Type 3 cities. They are most dissimilar from Type 1 and Type 5 cities, especially in terms of DIFF.

4. Discussion

4.1. Discussion of types

4.1.1. The large cities: very low vs. high CONS (Types 1 & 4)

The largest cities, with few exceptions, appear in Type 1 and Type 4. The main distinction between Type 1 and Type 4 is that Type 1 cities have very low CONS while Type 4 cities have high CONS. As would be expected, Type 1 cities have a WUCI that is much lower than that of Type 4 cities. However, Type 1 cities tend to have a low DIFF, and this means that the WUCI of Type 1 cities is more similar to that of Type 2 and Type 3 cities than it would be otherwise.

The high per capita water consumption exhibited by Type 4 cities suggests that these cities currently have sufficient water supply and urban water infrastructure. Type 4 cities also tend to be located in areas with higher natural water availability than several other city types, such as Type 1. However, because the population of Type 4 cities tends to be so large and consumption is so high, these cities likely face many challenges in securing sufficient water supply now and in the future. This is indeed what we find. Even cities with abundant natural resources, such as Singapore and New York, have recently made substantial investments in promoting water conservation and in innovative water infrastructure (e.g., desalination and reclamation).

Both Type 1 and Type 4 cities are likely to have challenges in obtaining sufficient water resources for their large urban populations. However, differences exist between these two types. Many Type 1 cities are located in developing countries and undergoing rapid urbanization. Their water resources challenges may be heightened by low natural water availability, rapid urbanization, and low access to financial resources. The low water consumption in these cities may also be associated with relatively low levels of infrastructure. Even though Type 1 cities currently consume less water than Type 4 cities, it would not necessarily be appropriate to say they are “more sustainable” than Type 4 cities. If Type 4 cities are able to manage resources within their watershed areas properly, they may be as sustainable relative to their water supply as Type 1 cities. As Type 1 cities continue to grow and develop, so too will the size of their catchment areas; Type 1 cities may eventually become Type 4 cities.

Type 1 cities have the opportunity to implement new types of technologies and pioneering water management practices, such as more decentralized, neighborhood level water treatment and reuse, rather than trying to copy the development of water supplies in Type 4 cities. This would be an excellent opportunity for Type 1 cities to collaborate with Type 4 cities for knowledge exchange.

4.1.2. The small, wet cities (Type 3)

Type 3 cities are characterized by low populations that have low- to medium- levels of per capita water consumption and medium- to high- natural water supply. These cities are likely to have the fewest issues with sustainable water supply of any of the types. Water supply stresses may be less acute for Type 3 cities. However, in spite of relatively low CONS and abundant water resources, many of the cities in Type 3 still face challenges with sustainable urban water management. These may include changes in water quality due to urbanization, aging infrastructure, and flooding.

4.1.3. The medium-sized cities: medium vs. high CONS (Types 2 & 6)

With a few exceptions, medium-sized cities (those in the 2nd and 3rd quartiles of POP) tend to fall into Types 2 and 6. Type 6 cities have higher naturally available water resources. Another distinction between the two types is that Type 6 cities have very high CONS, which raises concerns about their sustainability, while those of Type 2 have medium-low CONS.

4.1.4. The arid cities: low DIFF (Type 5)

Type 5 cities were located in highly arid environments: these cities had negative values of DIFF, which means that there is much less rainfall than potential evapotranspiration. Yet surprisingly these cities were also distinguished by their high levels of water consumption (CONS). These cities are therefore likely to rely extensively on water transfers or water imports for water supply, either as rivers or canals (e.g., Cairo and Phoenix), conversion of salt water to freshwater (e.g., Dubai and Abu Dhabi), and mining of

fossil aquifers. Transporting water over large distances, desalination, and reclamation are all associated with relatively large costs and issues of financial sustainability and management capacity (Plappally & Lienhard, 2012; Plappally & Lienhard, 2013). The water resources for cities in Type 5 are particularly vulnerable to economic shocks, energy shortages, and political changes (since they may be obtaining water from outside their jurisdiction). Cost recovery and conservation may help Type 5 cities increase the resilience of their urban water systems to climate change and external pressures on water supplies.

4.2. Discussion of thresholds

The thresholds for DIFF suggest that cities in regions with naturally abundant water resources *may* be fundamentally different from those in more arid contexts. Types 3, 4, and 6 predominantly have values for DIFF *greater than zero*, which suggests that, all other things being equal, these cities may have relatively more ease at obtaining water resources than the other types.

The thresholds for POP raise intriguing questions. Are there significant differences in providing urban water resources for cities larger than 1 million? Types 1 and 4 tend to be the largest cities in the database, but these two types differ in CONS and DIFF. Cities of Type 1 have very low CONS and low DIFF, while those of Type 4 have high CONS and a positive DIFF. This suggests that city size is not deterministically related to climate or consumption.

4.3. Discussion of outliers and overlaps

It is important to remember that these types are based on the statistical clustering of continuous variables. Cities within a type are most likely to be more similar to each other than to other cities, and distinct from other types in some statistically significant way. The different types of cities identified in the clustering were relatively distinct on one or two variables but may have otherwise overlapped with another type on another. Because of this, cities with different membership may have similar metrics and therefore face some common sustainability challenges. For instance, Cairo and Paris are members of Type 5 and Type 2, respectively, and due to their relative sizes are likely to share similar challenges with cities of Type 1 and Type 4, respectively.

These overlaps and distinctions can be understood from Fig. 7, the scatterplot of the t-SNE results, in which the cities that fall close to one another are most similar, while those that are further apart are most dissimilar. Two cities that lie in a transition zone between adjacent types may be have different typological membership but be statistically similar to each other. Within the variable space shown in Fig. 7, it is also possible to visually identify sub-groupings. A subgrouping of Denver, Tbilisi, Vladivostock, Victoria, Stockholm, Dublin, Ottowa, and Kathmandu can also be identified in the dendrogram. Within this subgrouping there are non-intuitive pairings in the terminal leaf nodes, such as Kathmandu with Dublin and Vladivostock with Denver, which may identify unexpected similarities.

4.4. Discussion of intuitive and unexpected results

Some of these results are surprising while others are more expected. For instance, it is not much of a surprise that cities in very arid climates were grouped together. However, prior to application of the clustering algorithm this was not a foregone conclusion, and the results distinguished significant differences in water consumption of arid cities. This highlights the ability of the approach to yield meaningful results. It is perhaps also not much of a surprise that the largest cities were grouped together, as they were in Type 1 and in Type 4. At the same time, some may find it surprising to see London, Los Angeles, Singapore, Sydney, and New York together in a cluster in light of their climatic differences. However, examination of contemporary water issues in these cities supports this grouping. All of these cities have reached a size where water is imported from long distances and desalination has been at least considered if not implemented; all of these cities have issues with water quality and stormwater runoff. Water sustainability plans for these cities may thus be expected to have some similarities. It might also have been expected that large, rapidly urbanizing cities in developing countries would be grouped together, as they were in Type 1. For instance, it might have been possible to identify São Paulo and Mumbai as having similar challenges in meeting the demands of their burgeoning populations, but we are not aware of previous research that has compared these urban water systems. It is notable that this meaningful pairing, and others, were identified through a relatively simple combination of metrics, with relatively widely available data, using a very scalable method.

The results also identified more surprising international groupings. Consider the subgrouping of Hanover, Durban, Copenhagen, Vilnius, and San Salvador in Type 3. While it is perhaps less surprising that Copenhagen, Hanover, and Vilnius are similar, it is surprising to include Durban and even more so San Salvador with those three. Yet upon examining the scatterplots of these mid-range values on most variables, it makes sense why these cities were clustered. The identification of such sub-groups demonstrates the utility of applying quantitative data-mining algorithms to uncover significant and intriguing patterns of cities around the world.

5. Conclusions and future work

Our results provide an initial answer to the question—how similar and how different *are* cities in terms of their water resource supply and usage patterns? One of the main aims of this study was to use simple metrics for water supply and demand to identify groups of similar and different cities. Using statistical clustering identified six meaningful clusters of urban water sustainability conditions. We recast these clusters as a typology, and indicated how our results can be used for stratified sampling of smaller numbers of cases for more fine-grained contextual and comparative international water research.

Our work also provides a context for assessing water management challenges and policy alternatives in a meaningful way. It identified both expected and surprising pairings and groupings of cities that can be explored through further comparative analysis in small-N or medium-N studies. For instance, the typology identified unanticipated similarities, most notably among cases of cities that consumed far more than their local supply, even when those supplies varied from arid to humid.

Our results primarily demonstrate the utility of statistical clustering of a small number of metrics to support meaningful international comparison of cities. The relevance of this typology to agenda-setting and policy-making for sustainability depends on the assertion that cities are more likely to share similar sustainability challenges with other cities that have similar supply and demand metrics. While we believe that cities within a type, or adjacent to each other in Fig. 7, are more likely to share similar issues in urban water resource management, the relative proximity or dissimilarity of two cities alone cannot determine whether the most pressing water resource and management challenges these cities face are, in fact, similar. Further work that links small-N, medium-N, and large-N scales of comparative analysis must be done to uncover more nuanced predictive relationships between water supply and demand metrics and sustainability issues.

The thresholds in our results provide a basis for posing questions with testable hypotheses. The thresholds for POP, CONS, and DIFF are intriguing, but whether they are in fact meaningful requires further study. For instance, the levels for these thresholds or typological membership of particular cities may change if the database of cities is expanded or if the clustering is repeated on the same set of cities at a different point in time. Hopefully, as time progresses we might see the emergence of new types, such as that of large, affluent cities with low water use.

A next step in this analysis would be to examine the sensitivity of the clustering to the underlying dataset. There are other opportunities for next steps in the analysis. High-priority extensions of this analysis are expanding the set of cities and integrating additional metrics, such as a water quality variable, partitioning water consumption into municipal and industrial components, and adding metrics associated with more socio-economic aspects of water management, including GDP or household income. Including data on the financial, energy, material, and land use intensity of urban water supply could provide insight into the urban water-land-energy nexus. The dataset could also be augmented by the inclusion of performance indicators such as leakage rates and cost recovery rates included in databases such as IBNET. This would allow for a more refined approach to identifying types of water systems and target areas to enhance the sustainability of urban water management.

Another attractive step would be to include spatial and temporal variations in the underlying data. For instance, the climate metric could be disaggregated to a monthly water budget. This would allow for distinctions to be made among cities with large intra-annual variation in their water budgets, since this variation can have important implications for water storage, management, and reuse. Expanding the dataset to include multiple years of data for cities might allow for identification of types based on variability and trends—for instance, cities that have increasing or decreasing per capita water use (CONS).

In summary, we applied statistical data-mining techniques to develop an initial urban water typology based on city size, per capita water consumption, and net annual water balance. This typology is an important first step in characterizing urban water supply and demand patterns around the world and will facilitate transfer of knowledge and meaningful case study research to further our understanding of sustainable urban water resources management. We demonstrated that statistical clustering is a useful method for developing a quantitative basis for small-N and large-N comparative urban water management case study research. The typology presented here is a significant contribution to this effort, but it is only a start. It will benefit from future case study research, standardization of data, expansion of metrics to consider temporal and spatial variation, disaggregation of water consumption and net annual water balance data, and the inclusion of more cities.

Acknowledgements

This research was made possible with generous funding from the Rambøll Foundation and Liveable Cities Lab for a research project on Enhancing Blue-Green Infrastructure and Social Performance in High Density Urban Environments (2014–2016). We thank Richard de Neufville, anonymous reviewers, and the editors for helpful feedback.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.scs.2016.06.003.

References

Astel, A., Tsakovski, S., Barbieri, P., & Simeonov, V. (2007). Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. *Water Research*, *41*(19), 4566–4578.

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., et al. (2012). Smart cities of the future. *The European Physical Journal: Special Topics*, *214*(1), 481–518.

Bettencourt, L. (2013). The origins of scaling in cities. *Science*, *340*(6139), 1438–1441.

Candelieri, A., & Archetti, F. (2014). Identifying typical urban water demand patterns for a reliable short-term forecasting—the ICeWater project approach, 16th Conference on Water Distribution System Analysis. *Procedia Engineering*, *89*, 1004–1012.

Chertow, M. R. (2001). The IPAT equation and its variants. *Journal of Industrial Ecology*, *4*(4), 13–29.

Daniels, P. L., Lenzen, M., & Kenway, S. J. (2011). The ins and outs of water use—a review of multi-region input-output analysis and water footprints for regional sustainability analysis and policy. *Economic Systems Research*, *23*(4), 353–370.

Diao, K., Farmani, R., Fu, G., Astaraie-Imani, M., Ward, S., & Butler, D. (2014). Clustering analysis of water distribution systems: identifying critical components and community impacts. *Water Science & Technology*, *70*(11), 1764–1773.

Ercin, A. E., Aldaya, M. M., & Hoekstra, A. Y. (2011). Corporate water footprint accounting and impact assessment: the case of the water footprint of a sugar-containing carbonated beverage. *Water Resources Management*, *24*, 721–741.

Ferrão, P., & Fernández, J. (2013). Urban typologies: prospects and indicators. In *Sustainable Urban Metabolism*. Cambridge, MA: The MIT Press (Chapter 7)

Fletcher, T., & Deletic, A. (Eds.). (2008). *Data requirements for integrated urban water management: urban water series-UNESCO-IHP* (Vol. 1). Paris, France: CRC Press.

Gandy, M. (2014). *The fabric of space: water, modernity, and the urban imagination*. Cambridge, MA: The MIT Press.

Garg, V. (2007). *Forecasting the water demand using regression and cluster analysis for Salt Lake City. Masters thesis*. Utah State University.

Hoekstra, A. Y., & Chapagain, A. K. (2007). Water footprints of nations: water use by people as a function of their consumption pattern. *Water Resources Management*, *21*(1), 35–48.

Hoff, H., Döll, P., Fader, M., Gerten, D., Hauser, S., & Siebert, S. (2014). Water footprints of citiesöindicators for sustainable consumption and production. *Hydrology and Earth System Sciences*, *18*, 213–226.

IBNET. (2015). *IB-NET Database*. https://database.ib-net.org/ Accessed 28.01.16

Kennedy, E. H. (2011). *Reclaiming consumption: sustainability, social networks, and urban context. Doctoral dissertation*. Rural Sociology, University of Alberta.

Khamis, A. (2012). Developing a typology of urban resource consumption. In *1st Civil and Environmental Engineering Student Conference*.

Kim, J.-S. (1997). *The differentiation of South Korean cities: a multivariate analysis. Master's thesis*. Department of Geography, The University of Calgary.

Konar, M., Dalin, C., Hanasaki, N., Rinaldo, A., & Rodriguez-Iturbe, I. (2012). Temporal dynamics of blue and green virtual water trade networks. *Water Resources Research*, *48*, W07509.

MacCannell, E. H. (1957). *An application of urban typology by cluster analysis to the ecology of ten American cities. Doctoral dissertation*. University of Washington.

Mayer, A., Winkler, R., & Fry, L. (2014). Classification of watersheds into integrated social and biophysical indicators with clustering analysis. *Ecological Indicators*, *45*, 340–349.

McDonald, R. I., Weber, K., Padowski, J., Flörke, M., Schneider, C., Green, P. A., et al. (2014). Water on an urban planet: urbanization and the reach of urban water infrastructure. *Global Environmental Change*, *27*(2014), 96–105.

Mollinga, P., & Gondhalekar, D. (2014). Finding structure in diversity: a stepwise small-N/medium-N qualitative comparative analysis approach for water resources management research. *Water Alternatives*, *7*(1), 178–198.

Novotny, V., Ahern, J., & Brown, P. (2010). *Water centric sustainable communities: planning, retrofitting and building the next urban environment*. New York, NY: Wiley.

Plappally, A. K., & Lienhard, J. H. (2012). Energy requirements for water production, treatment, end use, reclamation, and disposal. *Renewable and Sustainable Energy Reviews*, *16*, 4818–4848.

Plappally, A. K., & Lienhard, J. H. (2013). Costs for water supply, treatment, end-use and reclamation. *Desalination and Water Treatment*, *41*(1–3), 200–232.

Rahill-Marier, B., & Lall, U. (2013). *America's water: an exploratory analysis of municipal water survey data. White paper*. New York: Columbia Water Center, Earth Institute, Columbia University. http://water.columbia.edu/2013/10/16/americas-water-an-exploratory-analysis-of-municipal-water-survey-data/ Accessed 14.12.15

Rao, A. R., & Srinivas, V. V. (2008a). Regionalization by hybrid cluster analysis. In *Regionalization of Watersheds*. pp. 17–55. Dordrecht, Netherlands: Springer (Chapter 2).

Rao, A. R., & Srinivas, V. V. (2008b). Regionalization by fuzzy cluster analysis. In *Regionalization of Watersheds*. pp. 57–111. Dordrecht, Netherlands: Springer (Chapter 3).

Rosa, E. A., York, R., & Dietz, T. (2004). Tracking the anthropogenic drivers of ecological impact. *Ambio*, *33*(8), 509–512.

Saldivar-Sali, A. N. D. (2010). *A global typology of cities: classification tree analysis of urban resource consumption. Master's thesis*. Department of Architecture, MIT.

Suweis, S., Konar, M., Dalin, C., Hanasaki, N., Rinaldo, A., & Rodriguez-Iturbe, I. (2011). Structure and controls of the global virtual water trade network. *Geophysical Research Letters*, *38*, L10403.

van der Maaten, L. J. P., & Hinton, G. E. (2008). Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Wörlen, M., Wanschura, B., Dreiseitl, H., Noiva, K., Wescoat, J., & Moldaschl, M. (Eds.). (2016). *Enhancing blue-green infrastructure and social performance in high density urban environments: summary document*. Uberlingen, Germany: Ramboll Liveable Cities Lab.

Wescoat, J. L. (2009). Comparative international water research. *Journal of Contemporary Water Research & Education*, *142*, 61–66.

Wescoat, J. L. (2014). Searching for comparative international water research: urban and rural water conservation research in India and the United States. *Water Alternatives*, *7*(1), 199–219.

Willmott, C. J., Rowe, C. M., & Mintz, Y. (1985). Climatology of the terrestrial seasonal water cycle? *International Journal of Climatology*, *5*(6), 589–606.

Wu, S., Lv, M., Dong, S., Wang, J., & Xu, H. (2012). Classification calculation on water consumption of urban water supply network based on clustering. In *2012 World Automaton Congress* (pp. 1–4).

Yu, N., & Chen, R. (2010). Research on the development of urban infrastructure in China based on cluster analysis. In *2010 International Conference on Management and Service Science* (pp. 1–5). Published by IEEE.

Yu, Y., et al. (2013). Cluster analysis for characterization of rainfalls and CSO behaviours in an urban drainage area of Tokyo. *Water Science & Technology*, *68*(3), 544–551.