

# **A Clustering Analysis of Energy and Water Consumption in US States from 1985 to 2015**

Evgenia Kapousouz<sup>a,\*</sup>, Abolfazl Seyrfar<sup>b,\*</sup>, Sybil Derrible<sup>b</sup>, Hossein Ataei<sup>b</sup>

<sup>a</sup> Department of Public Administration, University of Illinois at Chicago

<sup>b</sup> Department of Civil, Materials, and Environmental Engineering, University of Illinois Chicago

\* These authors contributed equally to this work.

## **Introduction**

The world has a limited amount of natural resources and, as the world population keeps growing, the effective management of these resources in a sustainable way is paramount. In addition, according to the United Nations report on World Population Prospects (United Nations, 2017), the world population grows by approximately 83 million people every year. In fact, it is expected to reach 9.8 billion by 2050 and 11.2 billion by 2100, which will likely further exacerbate our current societal needs for additional resources in a resource-constrained world.

The proper management of energy and water are critical for a sustainable society (Derrible, 2017, 2018). Similar to most high-income countries, the overall demand for energy and water in the United States (US) has been increasing, partly as a result of population growth and changing climate conditions. To make matters worse, it is estimated that more than 50 percent of the US population is vulnerable to water resources risks (Padowski & Jawitz, 2012), and this problem is likely to become more severe in the future.

Both energy and water use and consumption are intrinsically interdependent, in what is commonly called the Energy-Water Nexus. For example, water is necessary to generate electricity and electricity is necessary for water withdrawal and distribution (i.e., to pump water) (Copeland & Carter 2017). In fact, most studies that analyze the Energy-Water Nexus describe it as a trade-off because large amounts of water are needed to generate electricity (Ruddell & Dixon, 2013). In terms of consumption, many studies examine water and energy consumption separately, but a gap exists in examining the relationship between the water and energy per capita consumption. In fact, evaluating water and energy consumption trends together may help us strengthen our understanding of how water and energy are used around the US (Ahmad and Derrible, 2015; Trabucco et al., 2019).

The main goal of this study is to leverage machine learning to cluster US states based on their similarities in terms of per capita water and energy consumption over time. The identification of these clusters can then shed light into understanding some of the trends and patterns occurring across the US. For this, we begin the chapter by explaining what sectors of energy and water consumption are considered to calculate the per capita values and what methods and techniques are used to analyze the data. We then analyze the data and illustrate the results so as to identify and further discuss the different clusters of states in five-year intervals from 1985 to 2015. We note, however, that justifying or explaining energy and water consumption similarities and differences between states is beyond the scope of this study. In other words, we identify clusters

and their evolution over time, but we do not explain why certain states are within the same cluster, which would require a study of the land use and socio-economic patterns between states.

## Materials and Methods

### *Data*

The data for water use were collected from the US Geological Survey (USGS, 2015), which is responsible for compiling water use data for the US at the county, state, and national level. The USGS publishes data every five years. It collects information on water use from local, state, and federal agencies for eight categories: (1) public supply; (2) domestic; (3) irrigation; (4) livestock; (5) aquaculture; (6) industrial; (7) mining; and (8) thermoelectric power. All 50 states were included in the data set, plus Washington DC. Technically, the data from the USGS is about water withdrawal (that includes leaks) as opposed to proper water use / consumption, but we will not make the distinction here and use the term “energy and water consumption.”

The current study uses the weighted domestic publicly- and self-supplied per capita use in liters per day (i.e., [L/d]). Public supply refers to water withdrawn by public and private water suppliers that provide water to at least 25 people or have a minimum of 15 connections. Self-supplied domestic water use is typically withdrawn from a private source or captured as rainwater in a cistern. It includes residential consumption (both indoor and outdoor) as well as water consumption from commercial and small industrial activities (USGS, 2015). At the national level, about 92% of the water was publicly supplied and 8% was self-supplied in 2015 (Dieter et al., 2018).

In this study, six years are included: 1985, 1990, 1995, 2005, 2010, and 2015. The year 2000 was not included because it did not contain the necessary information. The data for years 1995 and 2005 provided the per capita use per state. For the years 1985, 1990, 2010, and 2015, the values were calculated manually. The current study uses the following variables to estimate the water consumption per capita: (1) “Total population of county, in thousands;” (2) “Domestic, self-supplied population, in thousands;” (3) “Domestic, self-supplied per capita use, in gallons/day;” (4) “Domestic, publicly supplied per capita use, in gallons/day;” and (5) “Public Supply, total population served, in thousands.”

The data were collected at the county level and were aggregated to the state level. Specifically, for county  $i$ , the total population served in public supply ( $P_{ps,i}$ ) was multiplied with the publicly supplied per capita use ( $C_{ps,i}$ ). Similarly, the domestic self-supplied population ( $P_{ss,i}$ ) was multiplied by domestic self-supplied per capita use ( $C_{ss,i}$ ) and the two consumption values were summed for each county. Mathematically, the total population  $P_{c,i}$  and total water use  $W_{c,i}$  for county  $i$  is defined as:

$$P_{c,i} = P_{ps,i} + P_{ss,i} \quad (1)$$

$$W_{c,i} = P_{ps,i} \cdot C_{ps,i} + P_{ss,i} \cdot C_{ss,i} \quad (2)$$

Then the population  $P_{s,i}$  of state  $i$  was estimated by adding the total population per county  $P_{c,j}$  belonging to state  $i$ , such that:

$$P_{s,i} = \sum_j P_{c,j} \quad (3)$$

Similarly, the total water use  $W_{s,i}$  of state  $i$  was estimated by adding county level consumption, such that:

$$W_{s,i} = \sum_{j \in i} W_{c,j} \quad (4)$$

Lastly, the sum of the state level water consumption was divided by the sum of the state population to get the per capita water use  $W_{p,i}$  of state  $i$  as:

$$W_{p,i} = \frac{W_{s,i}}{P_{s,i}} \sum_{j \in i} W_{c,j} \quad (5)$$

The data for energy consumption was collected from the US Energy Information Administration (EIA 2019). The State Energy Data System (SEDS) gathers yearly data that are published by the EIA. The State Energy Data System (SEDS) collects data from primary and secondary sources of energy. Primary sources of energy like natural gas, coal, and petroleum are consumed directly whereas secondary sources of energy, such as electricity, are created by primary sources. The energy data used in the study include primary sources not destined to electricity (e.g., natural gas) and electricity from the secondary sources for the years when water consumption data were also available. The variables used are: (1) commercial sector consumption per capita and (2) residential sector consumption per capita.

In this work, we only consider energy consumption in the residential and commercial sectors so that energy consumption data used are consistent with the water use data. To determine how much energy,  $E_i$ , is consumed by a state  $i$ , we sum the energy consumption by the residential sector  $E_{r,i}$  in per capita megawatt-hours per year (i.e., [MWh/y]), and by the commercial sector  $E_{c,i}$  in per capita megawatt-hours per year as well (i.e., [MWh/y]), as shown below:

$$E_i = E_{r,i} + E_{c,i} \quad (6)$$

In addition, the original data collected from the US Energy Information Administration were in million British thermal unit (million Btu), and we used a thermal conversion factor of 0.293297 million BTU per megawatt-hour to convert the energy data to the metric system.

### *Climate Standardization*

To evaluate the effect of climate on energy use, we used heating degree days (HDD) and cooling degree days (CDD) to standardize energy use in each state. HDD captures the need for energy to heat buildings. CDD captures the need for cooling and air conditioning. Both HDD and CDD are calculated relative to a base temperature—that is, the temperature below/above which a building needs heating/cooling.

HDD and CDD data for all states was collected from the National Oceanic & Atmospheric Administration (NOAA), the National Weather Service, and the Climate Prediction Center at [ftp://ftp.cpc.ncep.noaa.gov/htdocs/degree\\_days/](ftp://ftp.cpc.ncep.noaa.gov/htdocs/degree_days/). Data for Alaska, Hawaii, and DC was not available, however, and we collected instead data for Anchorage, Honolulu, and Dulles International airport stations from the Western Regional Climate Center (WRCC) at <https://wrcc.dri.edu>. The base temperature for all HDD and CDD from these sources is 65° F.

To calculate the standardized energy use  $E_{cs,i}$ , we divide the entire yearly energy use by the sum of the HDD and CDD, such that:

$$E_{cs,i} = \frac{E_i}{HDD+CDD} \quad (7)$$

### Clustering Analysis

First, in terms of pre-processing, all water and energy consumption values are rescaled to fit in the same range, which is desirable to improve the clustering results. For this, we use the min-max scaling method because it maintains the shape of data distribution and it does not undermine the importance of outliers. With the min-max scaling method, the final standardized value  $X_s$  is defined as:

$$X_s = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (8)$$

where  $X$  is the initial value,  $X_{min}$  is the minimum value of the dataset for that year, and  $X_{max}$  is the maximum value of the dataset.

In machine learning, clustering is an unsupervised learning process to group a set of data objects into multiple groups, or *clusters*, based on similarity. In other words, data objects that belong to the same cluster should be relatively similar to one another and different from data objects in other clusters (Han, Pei, and Kamber, 2011). After considering several algorithms, in this study, we use hierarchical clustering to organize states based on their water and energy use values, essentially grouping the states that have similar use patterns together into clusters.

Hierarchical clustering can be either distance-based or density-based, and it can also adopt either a bottom-up (agglomerative) or a top-down (divisive) approach. In this work, agglomerative hierarchical clustering was used with the ward linkage algorithm, which is a distance-based algorithm that places the closest objects in the same cluster.

In agglomerative hierarchical clustering, each sample is in its own cluster at the beginning, and clusters are successively merged together. The ward linkage algorithm essentially minimizes the total within-cluster variance. To implement this method, at each step, we find the pair of clusters that leads to a minimum increase in total within-cluster variance after merging. This increase is a weighted squared distance between cluster centers. In the first step, each sample is one cluster, and the initial cluster distances are the squared Euclidean distance between points:

$$dist_{ward}(C_i, C_j) = d(i, j) = |X_i - X_j|^2 \quad (9)$$

To implement Ward's method, we can use the Lance–Williams algorithm. Lance and Williams (1967) established a recursive formula for updating cluster distances at each step when a pair of clusters is merged. For this, we need to find the optimal pair of clusters to be merged in each step, which is achieved by using the recursive formula. Assuming that clusters  $i$  and  $j$  are agglomerated into cluster  $i \cup j$ , and the dissimilarity  $d$  relative to an external cluster  $k$  is defined as:

$$d(i \cup j, k) = a(i) \cdot d(i, k) + a(j) \cdot d(j, k) + b \cdot d(i, j) + c \cdot |d(i, k) - d(j, k)| \quad (10)$$

where  $a(i)$ ,  $a(j)$ ,  $b$ , and  $c$  are parameters, which may depend on cluster size, that together with the cluster distance function  $d(i, j)$  form the algorithm. We can use Lance and Williams's formula as a

repeatedly executed recurrence and merge the clusters until reaching to our predetermined number of clusters.

Furthermore, the number of clusters selected by year is an exogenous parameter by the user, and it can differ from year to year. To determine the optimal number of clusters for each year, we used both the silhouette coefficient and visual inspection.

The silhouette coefficient is a similarity metric between objects in a dataset. Basically, the silhouette coefficient determines how well clusters are separated and how compact they are (Han, Pei, and Kamber, 2011). Suppose a dataset  $D$  divided into  $k$  clusters,  $C_1, \dots, C_k$ , for each object  $o \in D$ ,  $a(o)$  is the average distance between  $o$  and all other objects in the same cluster, and  $b(o)$  is the minimum average distance of  $o$  to all points in any other cluster. Based on the definition for  $a(o)$  and  $b(o)$ , for  $o \in C_i$  ( $1 \leq i \leq k$ ) we have:

$$a(o) = \frac{\sum_{o' \in C_i, o' \neq o} \text{dist}(o, o')}{|C_i| - 1} \quad (11)$$

$$b(o) = \min_{C_j: 1 \leq j \leq k, j \neq i} \left\{ \frac{\sum_{o' \in C_j} \text{dist}(o, o')}{|C_j|} \right\} \quad (12)$$

Here,  $a(o)$  captures the compactness of the cluster to which  $o$  belongs. The smaller  $a(o)$  is, the more compacted the cluster is, while  $b(o)$  reflects the dissimilarity of  $o$  with other clusters. The larger the value of  $b(o)$  is, the more separated  $o$  is from other clusters. Finally, by having the equations (9) and (10), we can calculate the silhouette coefficient of  $o$  as:

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}} \quad (13)$$

The silhouette coefficient value ranges from -1 to +1. If  $s(o)$  is closer to 1, then the cluster to which  $o$  belongs is more compact and  $o$  is more separated from other clusters. However, if  $s(o)$  is closer to -1, then  $o$  is closer to the objects in another cluster than to the objects in its cluster, which is not desirable (Han, Pei, and Kamber, 2011). Then, we can calculate the average silhouette coefficient of all data points in the dataset to assess the quality of the overall clustering process.

I The analysis was carried out in the Python language (Python 3.7) using the open source library Scikit-learn (Pedregosa et al. 2011).

## Maps

Maps were also produced to offer a visual representation of the results using ESRI's ArcGIS software package. ArcGIS is a platform used to share spatial data and make maps. The current study includes four maps for each year: (1) per capita water use by state, (2) per capita energy use by state, (3) clustering analysis results that considered both per capita water and energy use, and (4) clustering for water and climate-standardized energy use. On the maps, the numbers of categories match the numbers of clusters found during the clustering analysis. Through ArcGIS, one can easily visualize how consumption has changed over the years. For visualization purposes, the size of Alaska was reduced and the size of Hawaii was increased; these two states were also moved in the map. In general, darker colors indicate higher consumption. Finally, Washington DC was included in the dataset; however, because of its size, it could not be represented in the maps.

## Results

Figure 1 shows the evolution of per capita energy and water consumption in the US from 1985 to 2015.

Regarding per capita water consumption, we can see a small decrease from 1985 to 1995, followed by a stable trend around 980 L/day between 1995 to 2005. After 2005, per capita water consumption started to decrease again, but more significantly, with 2015 values being nearly 20% lower than 2005 values. This is in line with findings from the literature, including with Dieter et al. (2018) that finds a decreasing trend in water consumption since 1970.

A more in-depth look at the trends for each sub-category of water use (not included here) shows a systematic decrease after 2005 for almost all types of water uses (except mining and fish-farming). Specifically, we can see significant decreases for the thermoelectric (not studied here) and municipal sectors. The decrease in the municipal sector is a result of both improvements in the supply of and decreases in the demand for water. In terms of supply, infrastructure improvements have led to a reduction in leaks, and in terms of demand, Mayer et al. (2016) suggest that improvement in appliances (e.g., low-flow faucets and showerheads, low-flush toilets, and efficient washing machines) have led to a decrease in demand.

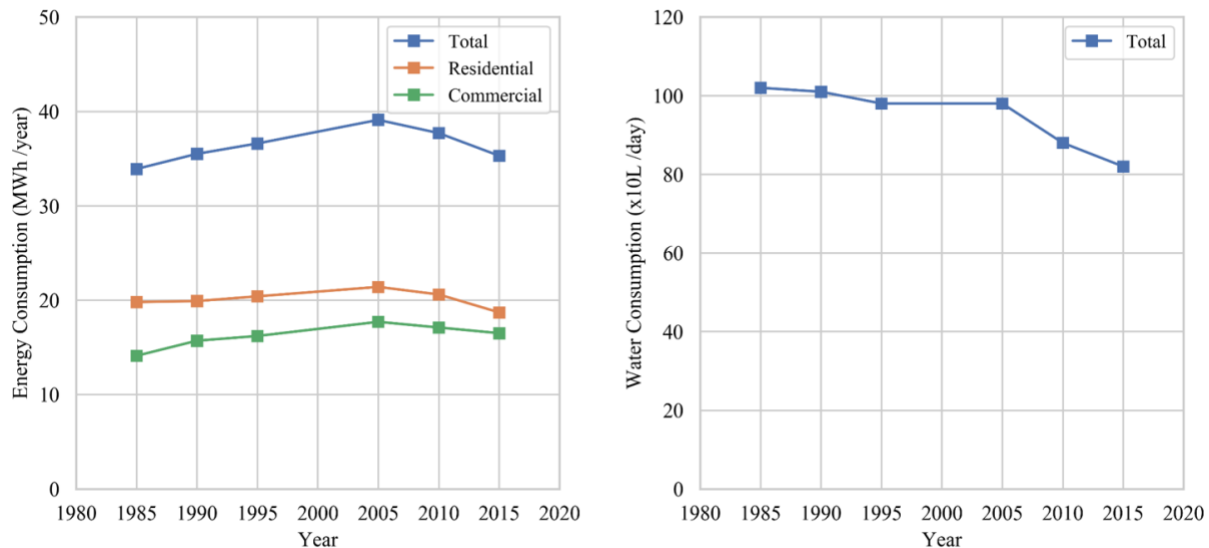


Figure 1. Evolution of Energy (Left) and Water (Right) Consumption in the US (Data Source: US Energy Information Administration, and United States Geological Survey)

Regarding per capita energy consumption in the residential and commercial sectors, we can see a steady increase in consumption from 1985 to 2005, followed by a steady decrease after 2005. This trend represents the average for the entire country, and not all states consume energy in the same way. In fact, since more than half of the energy use in homes tends to be for space conditioning (i.e., heating and cooling) (Energy Information Administration, 2019), geographic location and

climate is generally an important factor affecting energy consumption. Moreover, whether buildings use mostly natural gas or electricity for space conditioning can significantly affect energy consumption trends—although we recall that in this study both electricity and gas are included in our measure of energy.

We can further analyze energy consumption trends by taking into account the impact of regional climate conditions. From the data collected from the EIA and NOAA, generally, we find that states with the largest percentage decrease in per capita residential electricity consumption were also those with the largest decrease in HDD in the winter. Nevertheless, the Energy Information Administration (2017) also finds that decreases in residential and commercial energy consumption are also attributed to improvements in energy efficiency (e.g., for air conditioning and lighting).

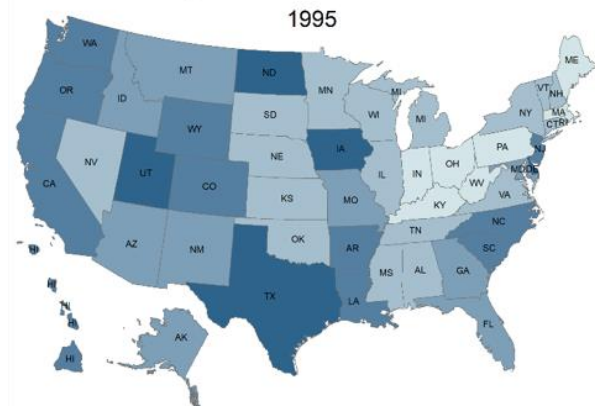
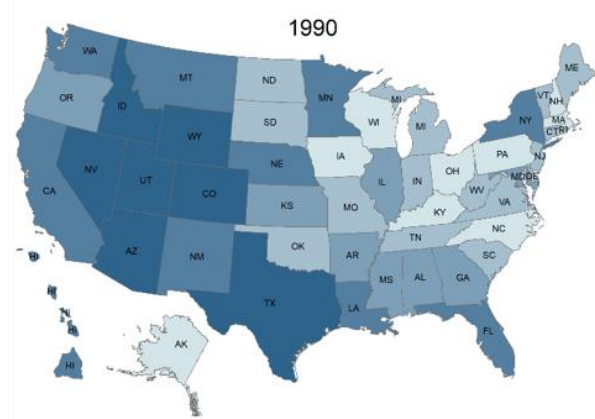
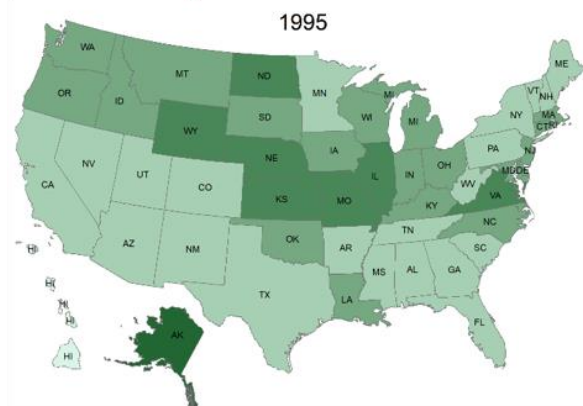
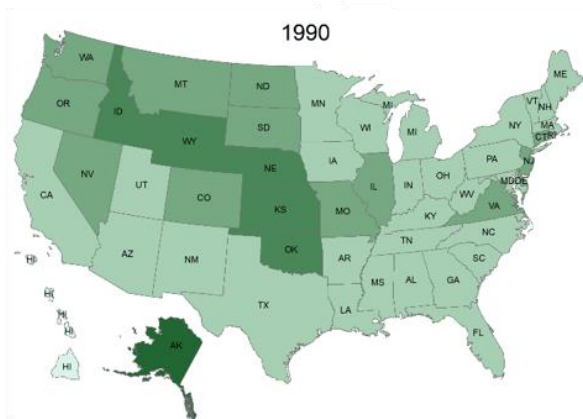
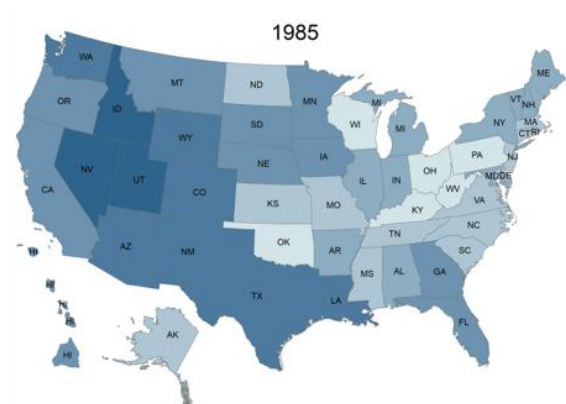
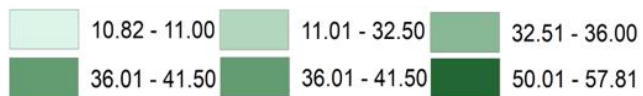
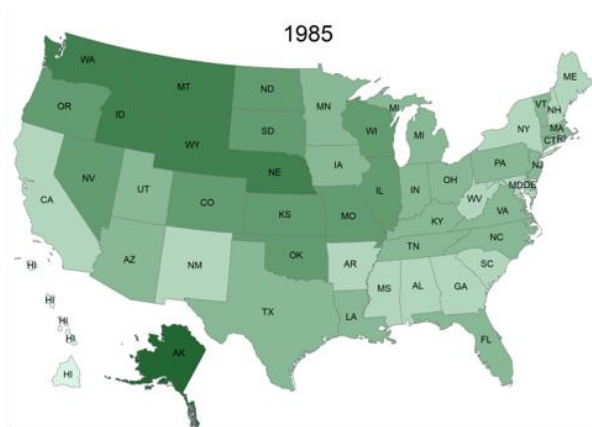
We can also study the evolution of energy consumption for every state from 1985 to 2015. Several states such as California, Hawaii, New York, and New Mexico consistently have the lowest per capita energy consumption values in the country. With the exception of 1985, Florida also has low values of energy consumption. Generally, states in the Southwest tend to have lower per capita energy consumption values compared to states in the Midwest. Moreover, several states such as Washington and Oregon actually reduced their per capita energy consumption over the entire period studied, unlike the average national trend that saw an increase from 1985 to 2005. Due to the extreme cold climate, Alaska is always among the states that consume the highest amount of energy, which is expected since building heating is energy-intensive. Similarly, northern states in mainland US tend to consume more energy than the national average. Out of all northern states, the per capita energy consumption in Wyoming and North Dakota were much higher than the others for the most years.

Generally, we find that neighboring states tend to have similar energy consumption trends. For example, Georgia, Alabama, and Mississippi have similar energy consumption trends over the time period studied, as do Wyoming, Nebraska, Kansas, Oklahoma, and Missouri. Moreover, states located in more temperate climates and states on the east coast like Florida, Georgia, Alabama, South Carolina, and Mississippi tend to be more stable and have only marginally evolved over time.

Moreover, we notice that states with large cities tend to have lower energy consumption values. For example, California and New York have constantly low energy consumption values. Illinois has low to medium energy consumption values. That being said, Texas follows somewhat different trends. In 1985 and 1990, Texas faired as a medium energy consumption states. In 1995, 2005, and 2010, energy consumption had decreased and had become “low,” before increasing again in 2015 such that Texas had become a medium energy consumption state again.

Figure 2 shows maps of per capita energy and water consumption for all years studied in this chapter.

Looking at per capita water consumption, it is apparent that although the clustering is different every year, permanent trends exist. The Northeast region (Maine, New Hampshire, Massachusetts, Rhode Island, New Jersey, Pennsylvania, Ohio, West Virginia, Virginia, West Virginia, Delaware, Maryland, and Kentucky) consistently show low to average water consumption trends. On the contrary, states in the Southwest region (Oregon, California, Nevada, Arizona, Utah, and Wyoming) achieve average to high water consumption trends.





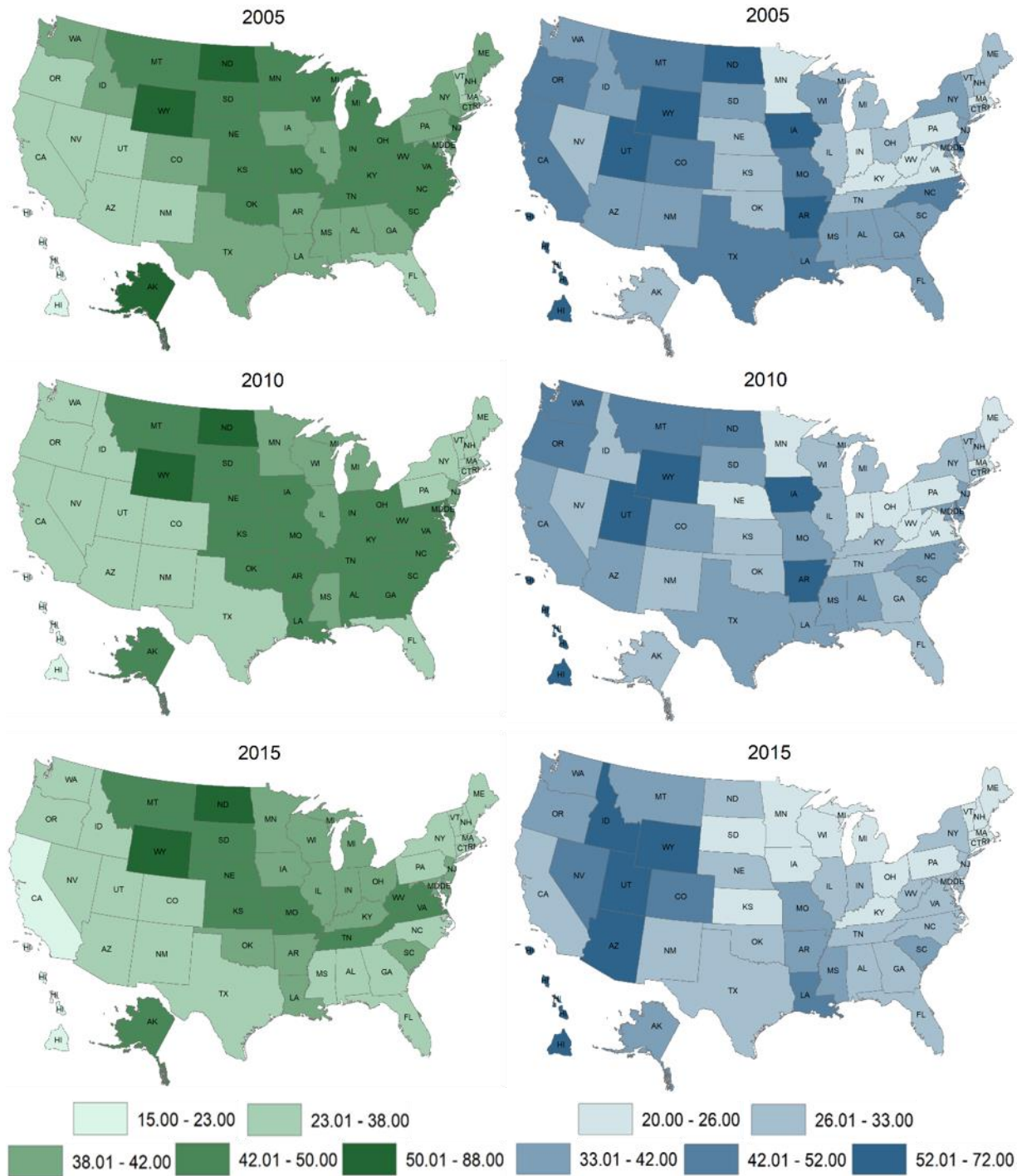


Figure 2. Energy Consumption (Left) versus Water Consumption (Right) by State from 1985 to 2015. Energy use is measured in MWh/year and water use is measured x10L/day.

This is in line with previous research from the US Environmental Agency (2004) that suggests that states in the West consume significantly more water. In addition, some states that show a gradual reduction in water consumption over the years, including Minnesota and Texas, while other states

such as Hawaii appear to have a slightly increasing trend over the years. Although not shown in the maps, Washington DC has been an outlier every single year with very high water consumption.

The clustering for water consumption has not changed significantly over the years for some of the states; for example, Florida is always classified as an average consumer. In contrast, some states have evolved significantly over the years. For example, water consumption in Iowa was classified as average in 1985, very low in 1990, very high in 1995, 2005, and 2010, and low in 2015; this phenomenon led us to investigate whether rainfall patterns had an impact on water consumption—little rain might entice people to water their lawns more—but we could not find any relationship. Similarly, Alaska's classification was not consistent; in 1985, it was classified as low consumption, in 1990 very low, in 1995 average, in 2005 and 2010 low, and in 2015 high. Furthermore, Nevada is classified as high water consumption in 1985 and 1990, water consumption then decreased in 1995, 2005, and 2010, but it increased again in 2015. No specific reasons were found to justify these patterns.

Finally, we find that some years (such as in 1990) have a high number of states classified as high water consumers, whereas other years (such as 2015) have a high number of states classified as low water consumers.

In the next step, the clustering analysis was done based on both water and energy consumption. The cluster size for each year was determined both by visual inspection and with the silhouette coefficient. As mentioned in the Materials and Methods section, the silhouette coefficient ranges from  $-1$  to  $+1$ , with values closer to  $+1$  indicating distinct clusters. As an example, Figure 3 shows the results for 5, 6, 7, and 8 clusters, and we can see that 6 clusters achieves the highest silhouette coefficient with a value of 0.65. In the end, it was found that a cluster size of five was found to be the best choice for all years except for the year 1985 when 6 clusters made more sense.

The results of the clustering analysis are illustrated in Figure 4. In 1985, although the per capita energy consumption of Alaska was higher than all other states, the cluster of High Water Consumption and High Energy Consumption is formed by DC and Idaho, and the cluster of High Water and Medium Energy Consumption was formed by the states of Nevada and Utah. The single outlier in this year is Hawaii, whose per capita energy consumption was almost 70 percent less than the national average.

In 1990, Idaho and DC significantly reduced their water consumption and joined to the Nevada and Utah cluster. Furthermore, DC increased its energy consumption, getting closer to Alaska that was the highest per capita energy consumer. Moreover, Hawaii was not an outlier anymore because its energy consumption had increased by about 70%, and with California, they now formed the MW\_LW (Medium Water Consumption and Low Energy Consumption) cluster.

Although DC significantly reduced its per capita water consumption throughout the time period, its per capita energy consumption has continued to increase, eventually becoming an outlier in 1995 with an energy consumption of more than 100% higher than the national average. Regarding the cluster HW\_HE in 1995, Idaho, Nevada, and DC were substituted by Iowa, North Dakota, and Delaware.

There were no substantial changes in terms of clustering patterns between 1995 and 2005 despite the 10-year gap. In 2005, DC remained an outlier because of its very high per capita energy consumption, and in the HW\_HE cluster, we have only one change as Wyoming was substituted

with Delaware. Similarly for 2010, the only change in the cluster HW\_HE was that Arizona joined to other four states. Finally, in 2015, DC was not an outlier anymore thanks to its reduction in energy consumption and the increase in energy consumption in North Dakota; both now formed a cluster.

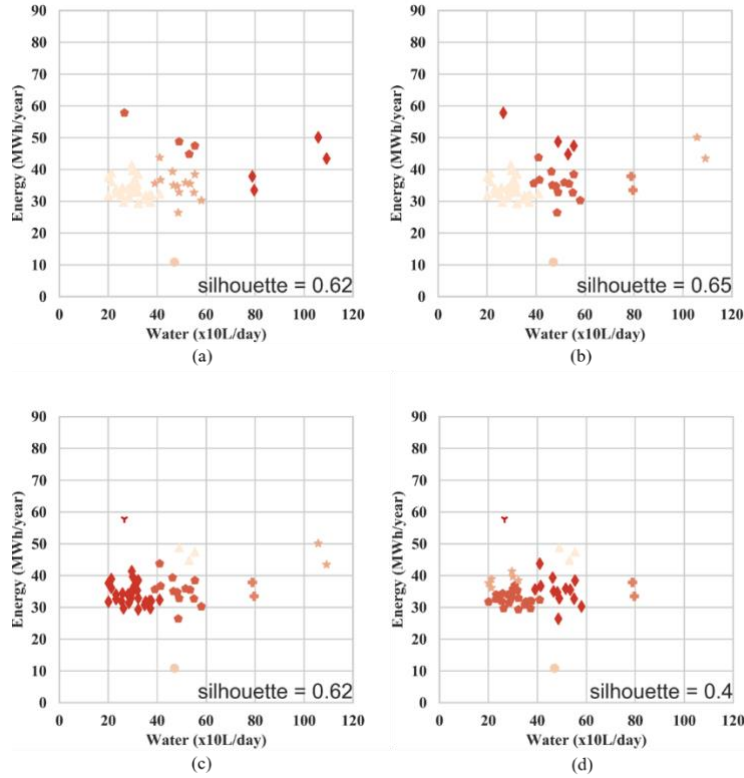
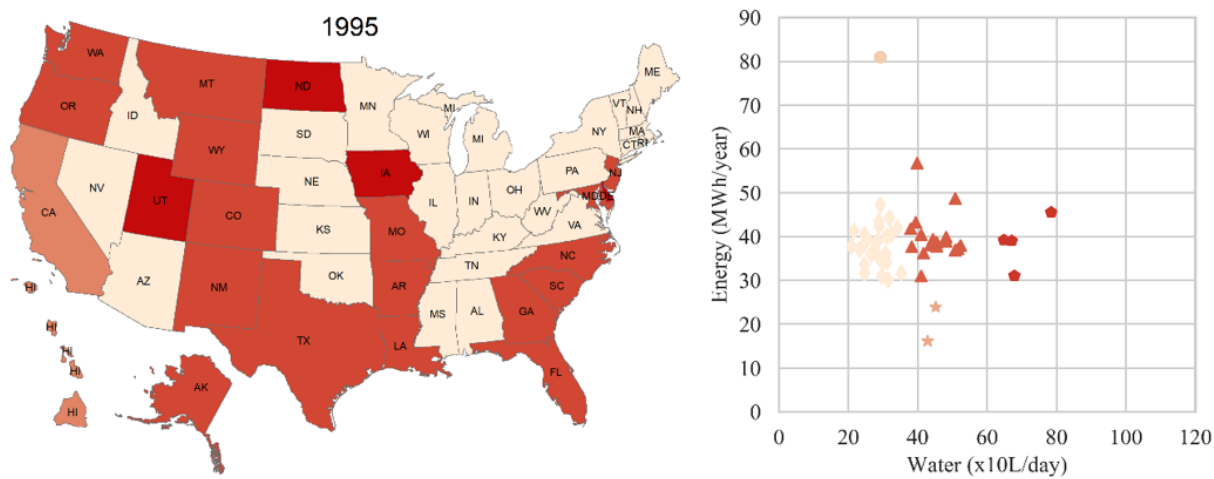
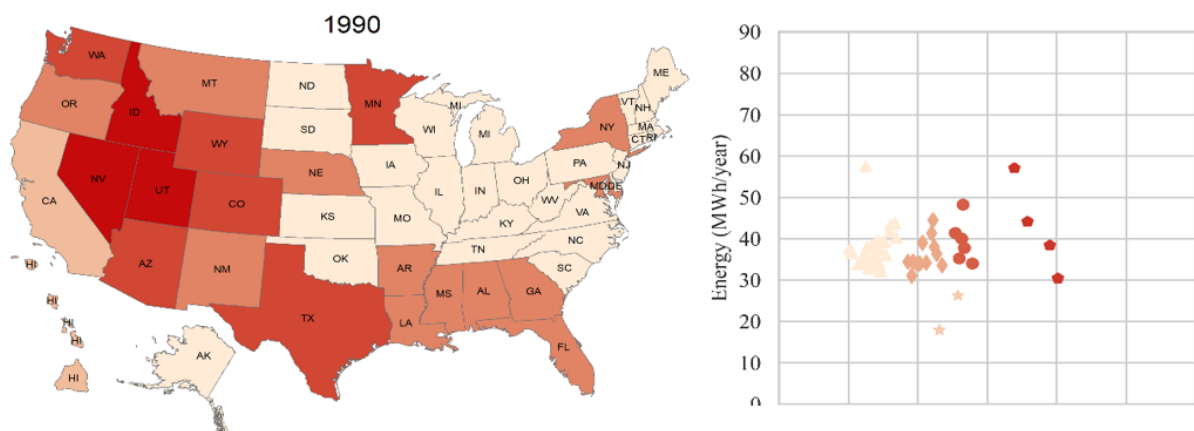
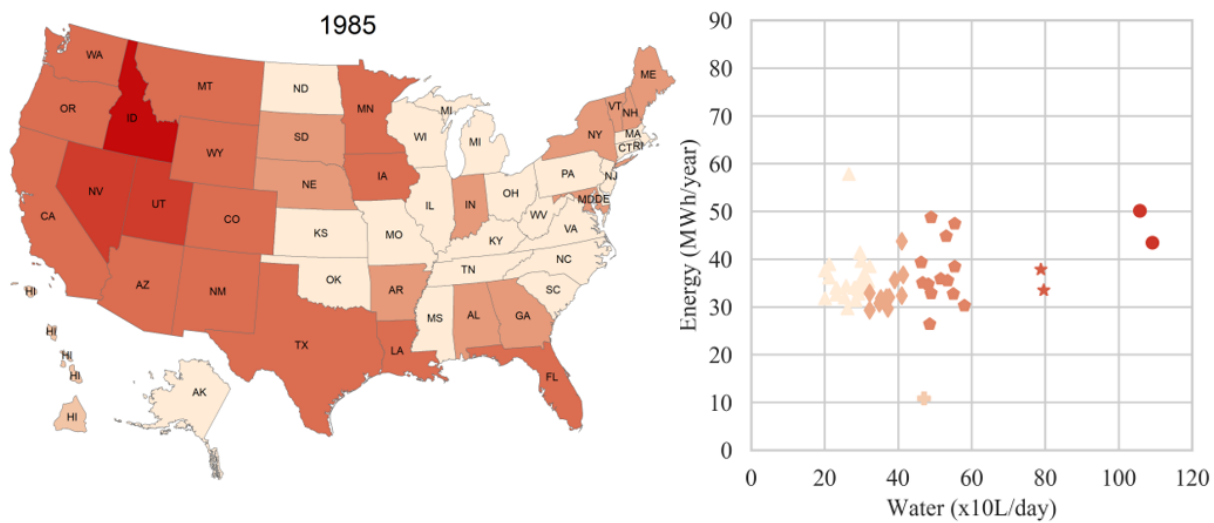


Figure 3. Illustrating the variation of Silhouette coefficient by changing the number of clusters:  
(a) 5 cluster, (b) 6 clusters, (c) 7 clusters, and (d) 8 clusters

As explained above, the silhouette coefficient determines how well clusters are separated and how compact they are. The value of the silhouette coefficient is between -1 and 1, and values closer 1 mean that the clusters are distinct and far away from one another, which is desirable. The silhouette coefficients calculated for all years ranged from 0.47 for 2005 to 0.71 for 1995, which is satisfactory. Some years show higher variability and clusters are more distinct (such as 1990), whereas other years have lower variability and clusters are not as distinct (such as 2010). For most years, the majority of the states are relatively close to each other despite the presence of some outliers. Interestingly, outliers actually vary over the years. For example, in 1985, Idaho and Washington DC were significantly high water consumers; in 1995, 2005, and 2010, Washington DC remained a significantly high energy consumer. In 1985, Hawaii was an outlier because it was a significantly low energy consumer compared to the other states.



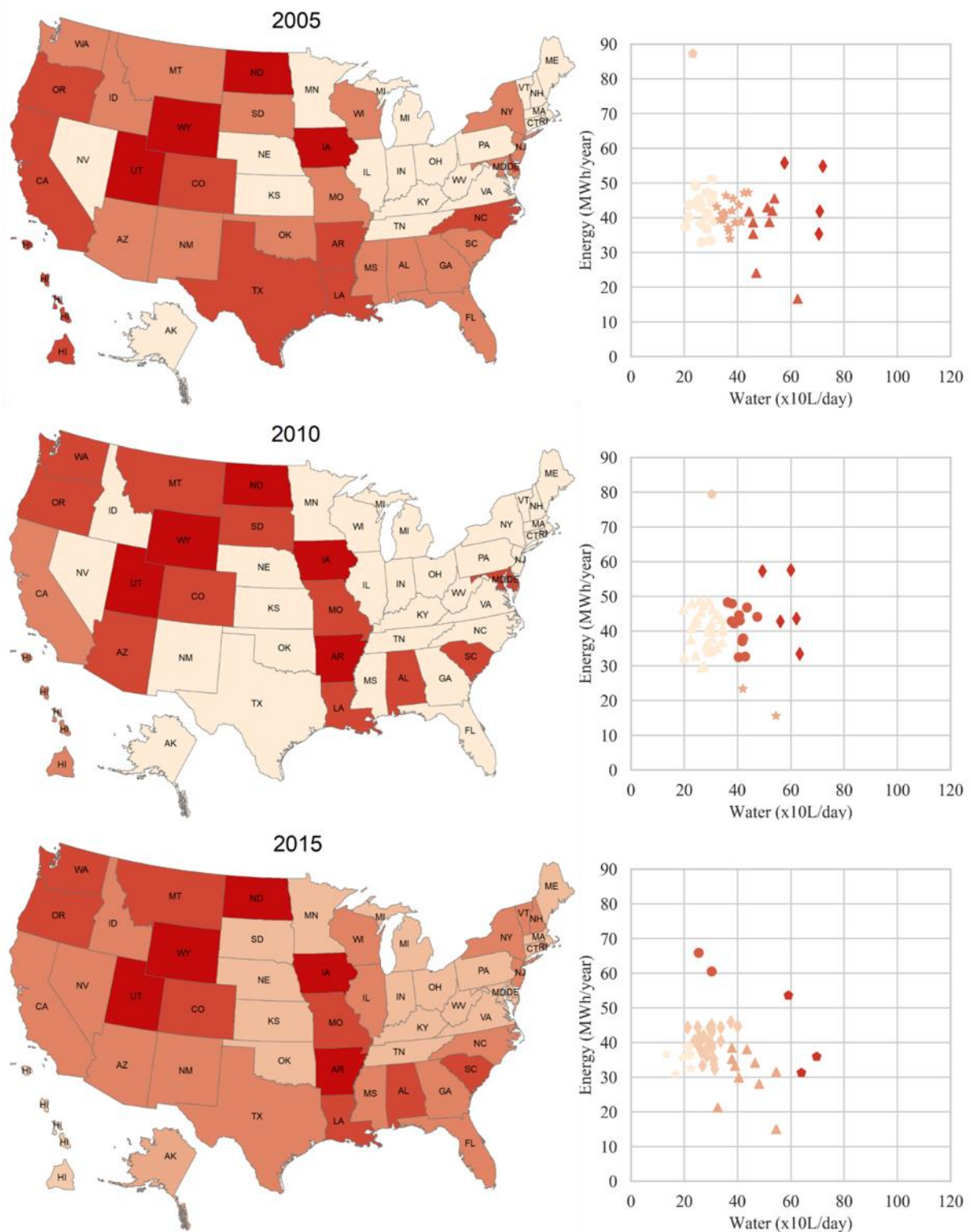


Figure 4. Clusters for Water and Energy Use by State from 1985 to 2015.

### *Evaluating the Impact of Climate on Energy Consumption Clustering*

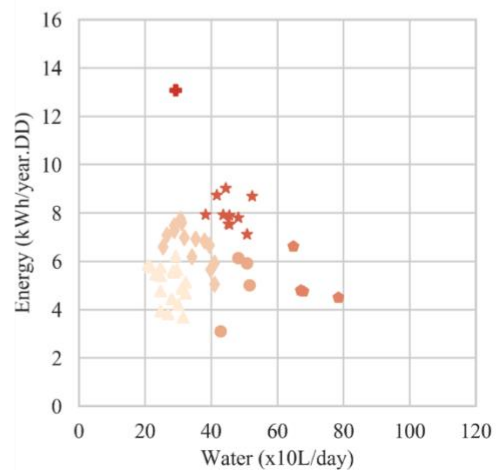
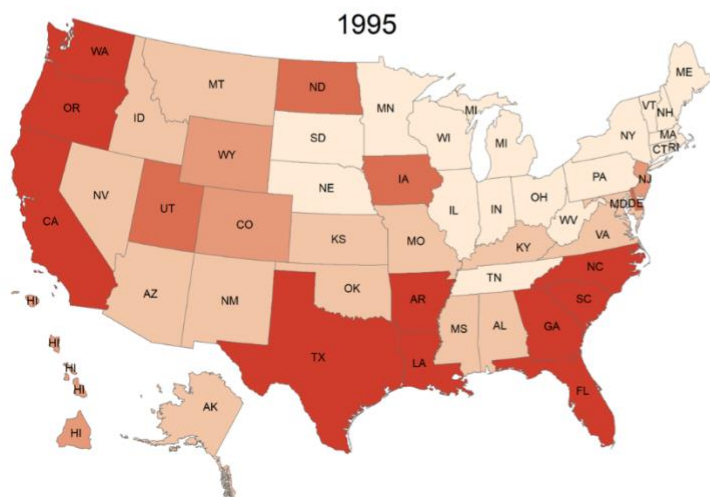
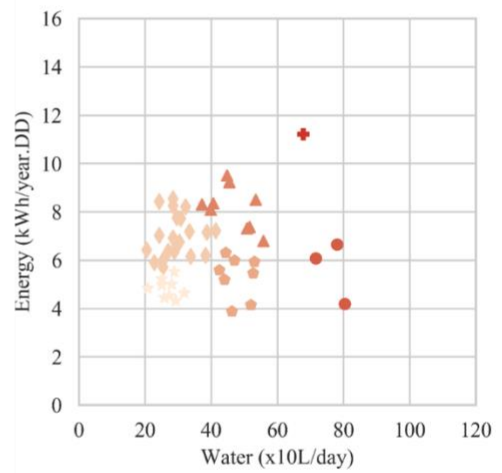
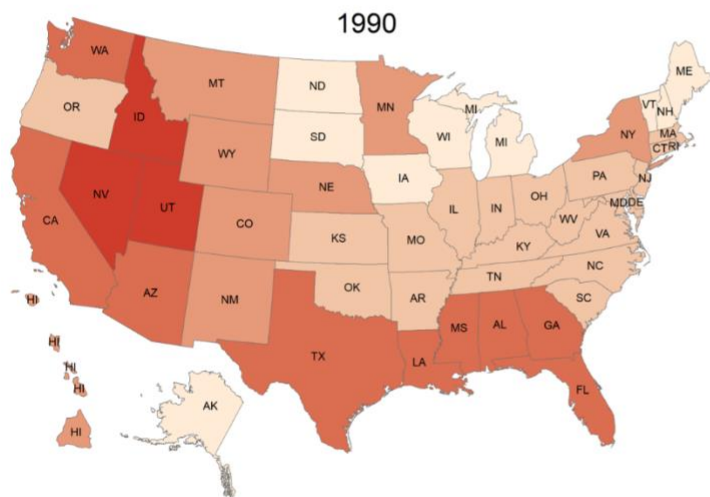
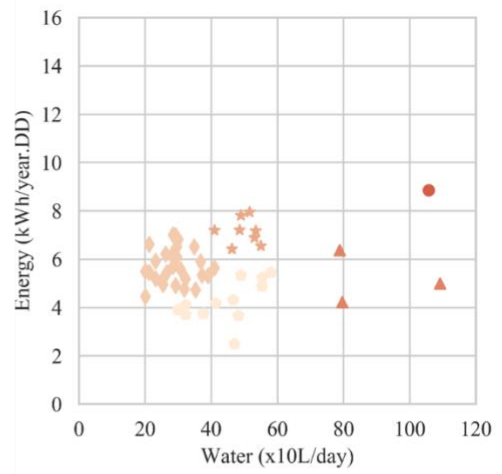
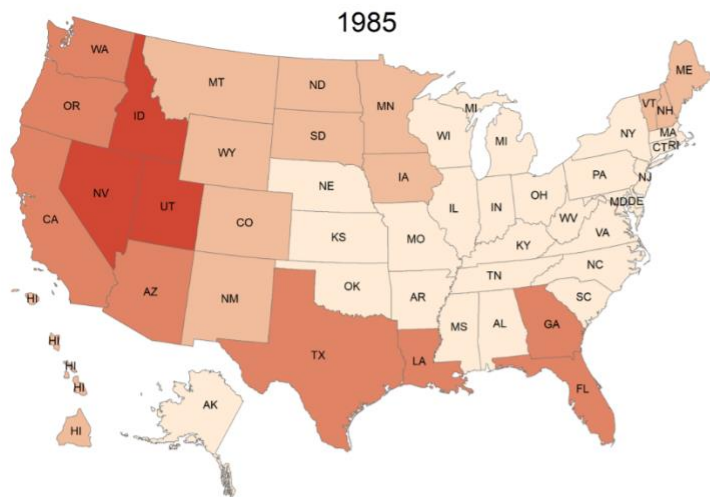
By using HDD and CDD to account for climate conditions, we standardized the yearly energy consumption patterns for each state and replicated Figure 4 with this new data. Figure 5 shows water and energy use by state, but using standardized values. Unlike the previous figures, Figure 5 does not have a legend because the range values selected to define the categories changed over the years, and including a legend of each map would clutter the figure and impact how the results would be visualized. With that in mind, the reader can identify the categories from the maps since the same colors were kept. The number of clusters changed after we standardized for climate, and we have 5 clusters for 1985 and 2010, and 6 clusters for 1990, 1995, 2005, and 2015.

After standardizing for climate, we can see several differences both in the clustering results and in the maps. In general, we can see that the distribution has changed, and the states seem more spread out on the maps. Moreover, by standardizing the data, we can see that Alaska is no longer an outlier, while DC remains an outlier. Moreover, Hawaii and California do not belong to the same clusters anymore. Generally, states in the Northeast exhibit lower energy and water consumption trends throughout the years. Water and energy consumption have not changed significantly over the years; however, the states do not always belong to the same cluster.

Specifically in 1985, Washington DC belonged to the very high water and energy consumption category, whereas Idaho, Nevada, and Utah belonged to the high water consumption category. All three states are close geographically, and historically Utah and Nevada have had similar population numbers between 1985 and 2015. In general, we can see that most states belong to the medium energy and water consumption category. In fact, water consumption had not changed significantly from 1985 to 1990, while energy consumption had increased. In addition, Washington DC has remained an outlier with high water and energy consumption—although the difference with other states is lower compared to 1990. Furthermore, Idaho, Nevada, and Utah have remained in the same cluster (i.e., high water consumption) albeit with lower values than in 1985. In addition, Utah remained in the same cluster, with high water consumption, to which Delaware, Iowa, and North Dakota were added in 1995.

Moving forward to 2005, water consumption had remained the same for most states, while energy consumption had slightly increased compared to 1995 values. This observation holds for DC, whose energy consumption further increased. In 2005, the states with high water consumption were Hawaii, Iowa, North Dakota, and Utah. By 2010, most states had similar water and energy consumption values. That being said, DC remained an outlier with very high energy consumption, although the states with the highest water consumption trends were Utah, Wyoming, Iowa, and Arkansas. Finally, by 2015, standardized energy consumption had slightly increased as well for most states. Energy consumption had decreased for DC, although it remained an outlier. States with higher water energy consumption were Hawaii, Arizona, Utah, Wyoming, and Idaho.





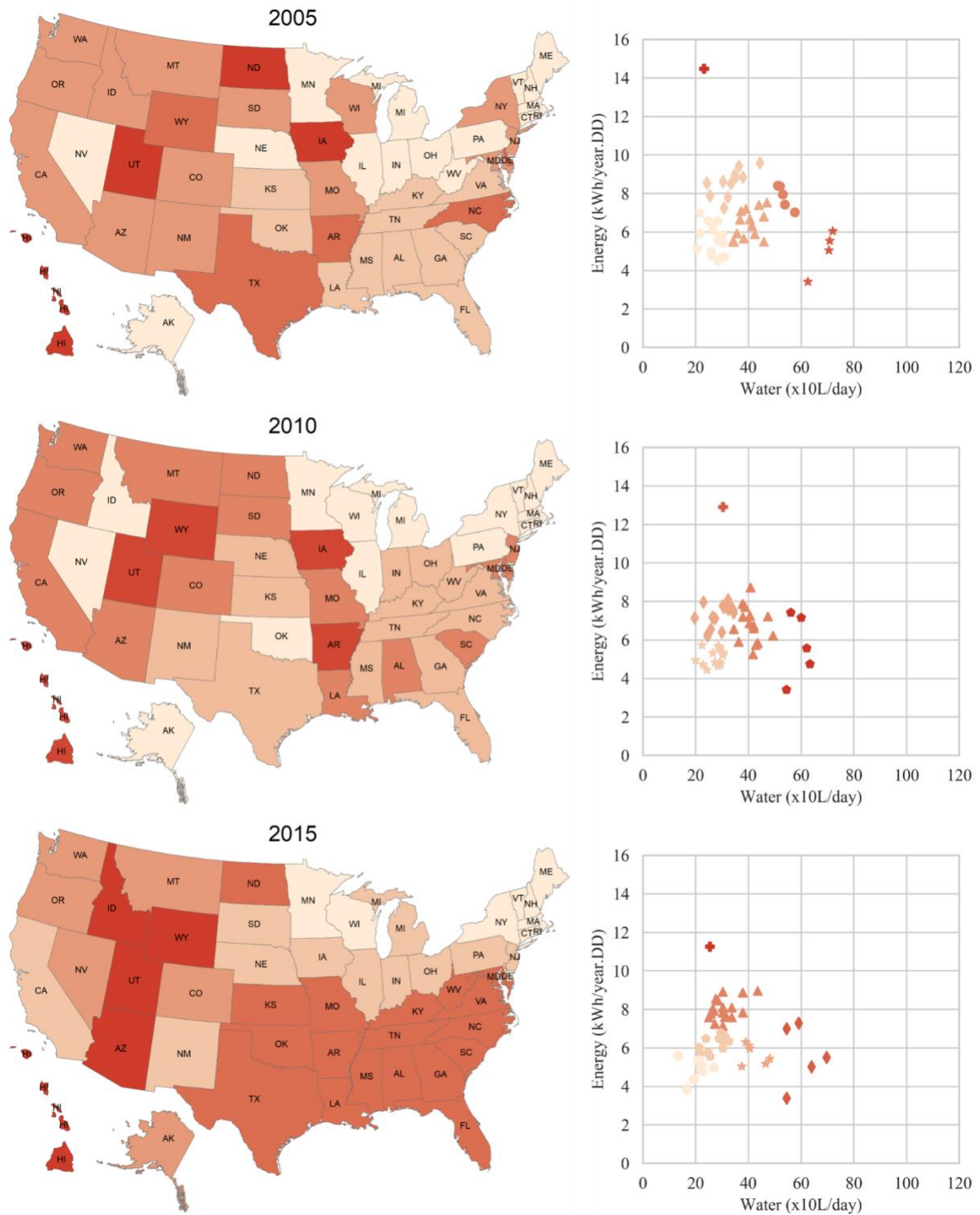


Figure 5. Clusters for Water and Weather Normalized Energy Use by State from 1985 to 2015.



## Conclusion and Discussion

In this chapter, we clustered all US states and DC from 1985 to 2015 based on their per capita water and energy consumption. The purpose of the study was to discuss trends and patterns over the time period—we did not attempt to fully explain or justify similarities and differences, which would require an in-depth study of the evolution of the land use and socio-economic characteristics of each state. We first looked at water and energy consumption separately and attempted to find clusters and trends over time. Then, by using clustering analysis, we took into account both energy and water consumption to identify different clusters of states over the 30-year time period.

Generally, we found a decreasing trend in per capita water consumption over the years. A contributing factor may be infrastructure as since 1985, many states have put significant efforts and investment into their water distribution systems, therefore reducing the amount of water lost through leakage. In contrast, per capita energy consumption had increased from 1985 to 2005, after which it started to decrease. Generally, we found that states in the Southwest tend to have lower per capita energy consumption compared to Midwest states that notably require more energy for building heating.

The results of the clustering analysis were evaluated through maps and clustering plots. To account for climate's significant impact on energy consumption, we also standardized energy consumption values by divided each value by the sum of the HDD and CDD for their respective states. We also tracked the significant changes in the clusters over the 30-year time period. For example, we noticed DC has reduced its per capita water consumption over the years and it has moved from the high consumer cluster in 1985 to the low water consumer cluster; however, its energy use has increased at the same time, eventually becoming an outlier.

Similarly, we examined changes in trends and patterns through the consumption and cluster maps and plots. Over the years, we noticed that states rarely remained in the same cluster. Two exceptions are California and Hawaii that stayed in the same cluster over the entire time period for the non-standardized clusters. By standardizing the data, we noticed that some neighboring states belonged in the same cluster for all years. Specifically, Indiana and Ohio belong to the same cluster, as do Georgia and Florida. In addition, Maine, Vermont, and New Hampshire remained in the same cluster throughout the 30-year period.

When examining per capita water and energy consumption separately, specific patterns became apparent. For example, states in the Northeast region tend to consume less water per capita, whereas states in the Southwest tend to consume less energy per capita.

The present study also possesses several limitations. Energy consumption data was collected annually, while water consumption data was collected every five years. A gap therefore exists to fully explore the trends between energy and water consumption. In addition, the data for 2000 was not available for water consumption, and therefore a gap exists between 1995 and 2005. Finally, the water data collected also had some limitations. In particular, the USGS does not account for water consumed with water bottles, which may skew some of the results and the consumption of water bottles has increased during the time period (International Bottled Water Association, 2019).

In the end, getting an appreciation for overall trends and patterns in energy and water consumption is paramount, and this chapter also demonstrated that data science techniques such as clustering can provide new and valuable insights.

## Acknowledgments

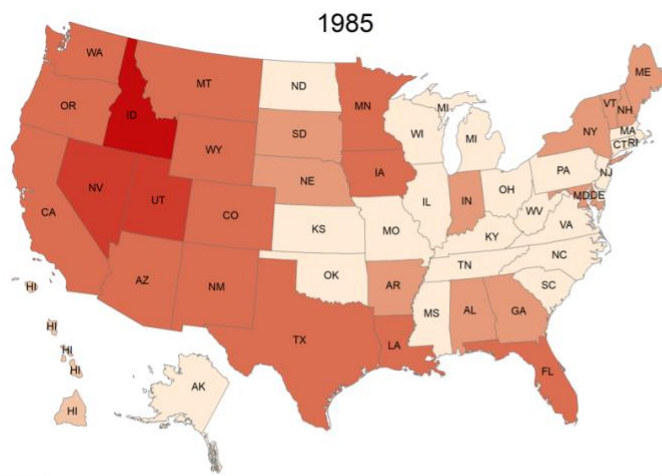
This research is partly supported by the National Science Foundation (NSF) CAREER award #155173.

## References

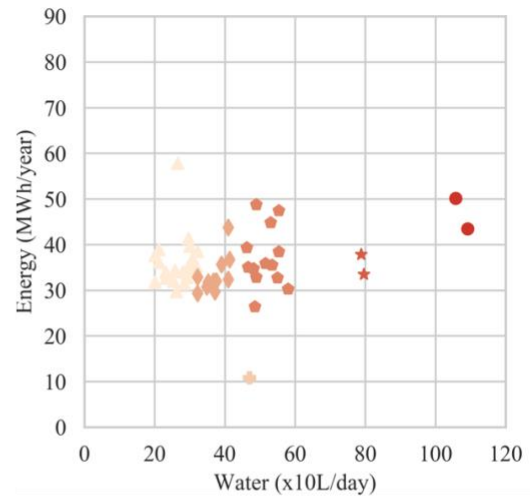
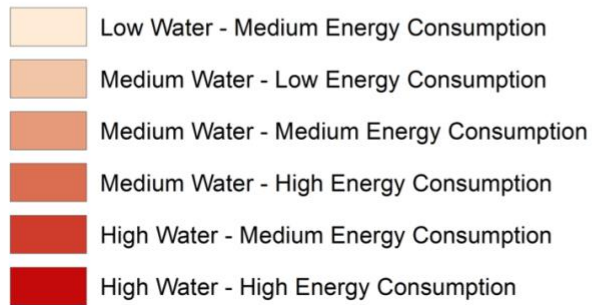
- Ahmad, N. and Derrible, S. (2015), Evolution of Public Supply Water Withdrawal in the USA: A Network Approach. *Journal of Industrial Ecology*, 19: 321-330.
- Balling, R. C., & Cubaque, H. C. (2009). Estimating future residential water consumption in Phoenix, Arizona based on simulated changes in climate. *Physical Geography*, 30(4), 308-323.
- Copeland, C. & Carter N. T. (2017). Energy-Water Nexus: The Water Sector's Energy Use. Library of Congress, Congressional Research Service.
- Derrible, S. (2017). Urban infrastructure is not a tree: Integrating and decentralizing urban infrastructure systems. *Environment and Planning B: Urban Analytics and City Science*, 44(3), 553–569.
- Derrible, S. (2018). An approach to designing sustainable urban infrastructure. *MRS Energy & Sustainability*, 5, E15.
- Dieter, C. A., Maupin, M. A., Caldwell, R. R., Harris, M. A., Ivahnenko, T. I., Lovelace, J. K., Harris, M. A., & Linsey, K. S. (2018). Estimated use of water in the United States in 2015. Circular. doi: 10.3133/cir1441
- EIA. State Energy Data System (SEDS): 1960-2016 (complete). US Energy Information Administration (2019). Available at <https://www.eia.gov/state/seds/seds-data-complete.php?sid=US>, accessed May 8, 2019.
- Energy Information Administration. Energy Use in Homes. (2019). Available at [https://www.eia.gov/energyexplained/index.php?page=us\\_energy\\_homes](https://www.eia.gov/energyexplained/index.php?page=us_energy_homes), accessed May 8, 2019
- Energy Information Administration. Today in Energy. (2017). Available at <https://www.eia.gov/todayinenergy/detail.php?id=32212>, accessed June 10, 2019
- Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.
- House-Peters, L. A., & Chang, H. (2011). Urban water demand modeling: Review of concepts, methods, and organizing principles. *Water Resources Research*, 47(5).
- International Bottled Water Association. (2019). Bottled Water – The Nation's Healthiest Beverage Sees Accelerated Growth and Consumption. Available at <https://www.bottledwater.org/bottled-water-%E2%80%93-nation%E2%80%99s-healthiest-beverage-%E2%80%93-sees-accelerated-growth-and-consumption>, accessed December 12, 2019.
- Iowa State University. Iowa Environmental Mesonet. Available at <https://mesonet.agron.iastate.edu/climodat/index.phtml?network=IAClimate&station=IA000&report=21>, accessed December 12, 2019.

- Lance, G. N., & Williams, W. T. (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems. *The computer journal*, 9(4), 373-380.
- Lee, D., & Derrible, S. (2019). Predicting Residential Water Demand with Machine-Based Statistical Learning. *Journal of Water Resources Planning and Management*, 146(1), 04019067.
- Mayer, P. W., DeOreo, W. B., Opitz, E. M., Kiefer, J. C., Davis, W. Y., Dziegielewski, B., & Nelson, J. O. Residential end uses of water. Available at [https://www.circleofblue.org/wp-content/uploads/2016/04/WRF\\_REU2016.pdf](https://www.circleofblue.org/wp-content/uploads/2016/04/WRF_REU2016.pdf), accessed December 12, 2019.
- Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?. *Journal of classification*, 31(3), 274-295.
- Padowski, J. C., & Jawitz, J. W. (2012). Water availability and vulnerability of 225 large cities in the United States. *Water Resources Research*, 48(12).
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel et al. 2011. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Ruddell, D. M., & Dixon, P. G. (2014). The energy–water nexus: are there tradeoffs between residential energy and water consumption in arid cities?. *International journal of biometeorology*, 58(7), 1421-1431.
- Trabucco, J.T., Lee, D., Derrible, S., Marai, G.E. (2019). Visual Analysis of a Smart City's Energy Consumption. *Multimodal Technologies Interact.* 3(30).
- United Nations. Department of Economic and Social Affairs, Population Division. World Population Prospects: The 2017 Revision, Key Findings and Advance Tables (2017). Available at <https://reliefweb.int/report/world/world-population-prospects-2017-revision-key-findings-and-advance-tables>, accessed December 12, 2019.
- U.S. Environmental Protection Agency. How We Use Water in These United States (2004). Available at <http://esa21.kennesaw.edu/activities/water-use/water-use-overview-epa.pdf>, accessed December 12, 2019.
- U.S. Geological Survey. USGS Water-Use Data Available from USGS (2017). Available at <https://water.usgs.gov/watuse/data/>, accessed December 12, 2019.
- World Commission on Environment and Development. (1987). Our Common Future. Oxford Paperbacks. Oxford; New York: Oxford University Press.

## Appendix








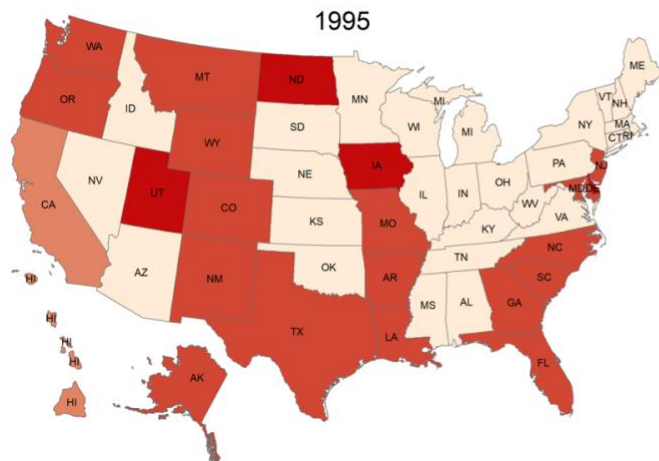
### Legend



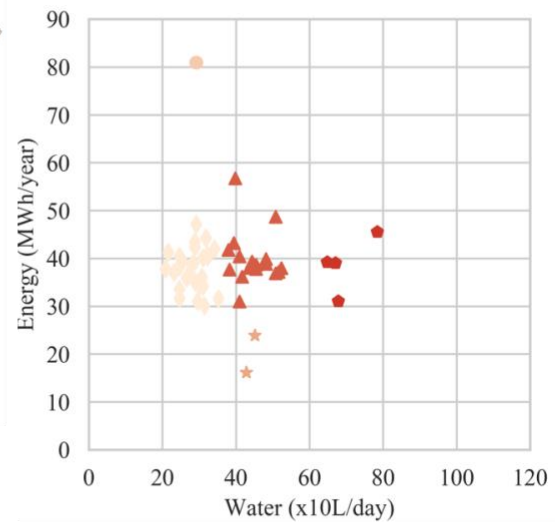
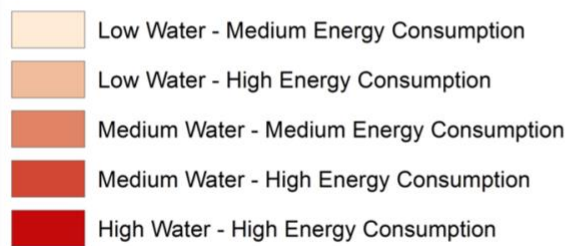


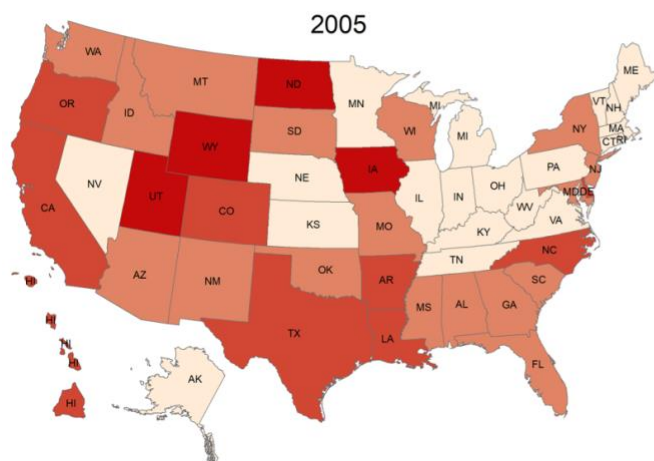
## Legend

- |   |  |
|---|--|
|  | Low Water - Medium Energy Consumption    |
|  | Low Water - High Energy Consumption      |
|  | Medium Water - Medium Energy Consumption |
|  | Medium Water - High Energy Consumption   |
|  | High Water - High Energy Consumption     |

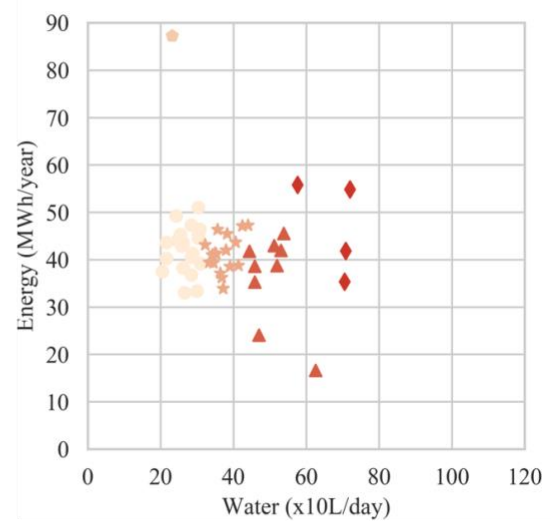
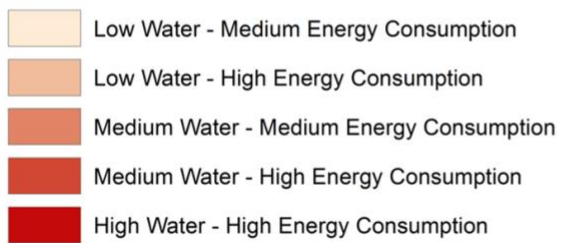


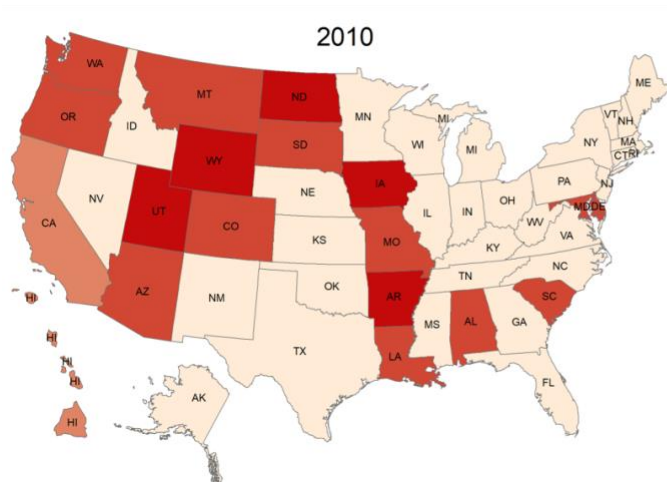
## Legend



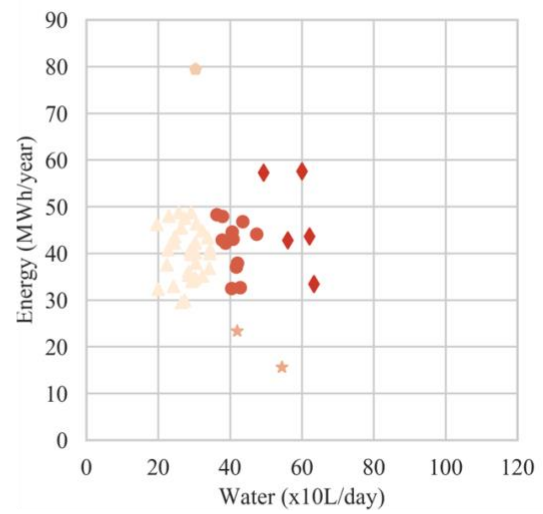
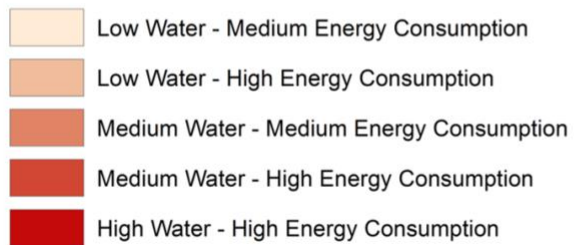


### Legend





### Legend





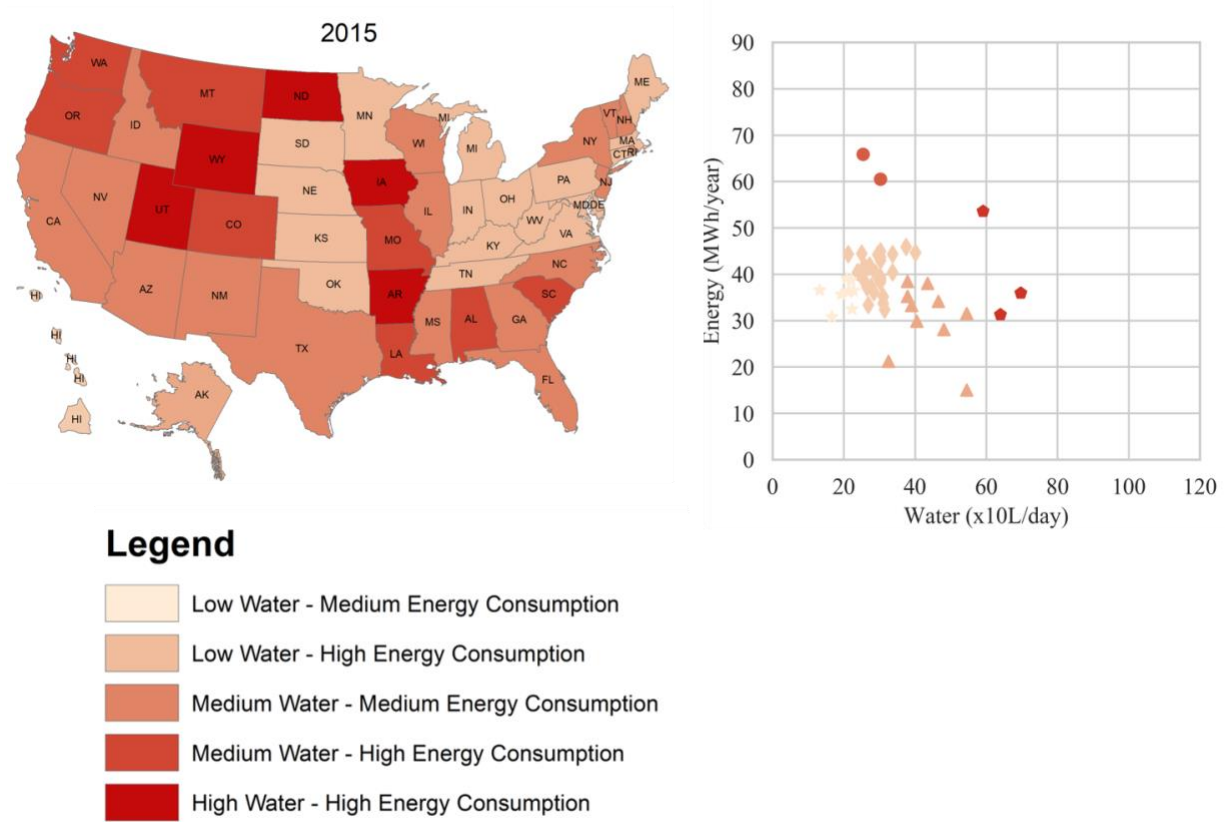
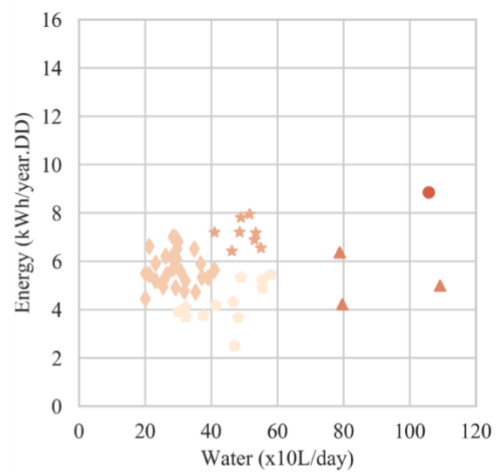
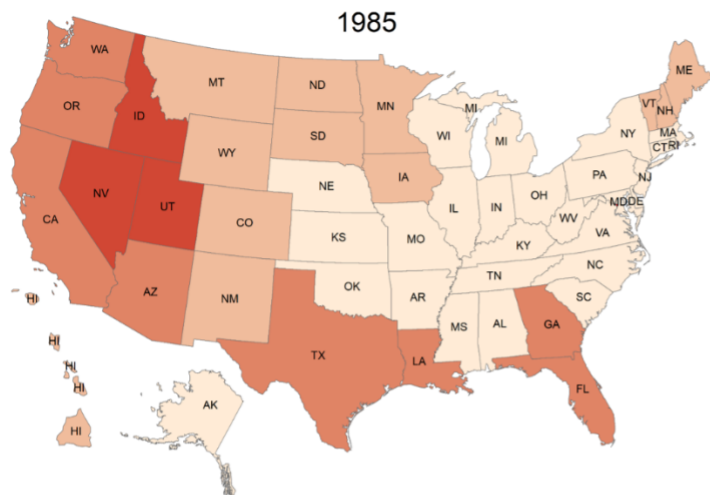
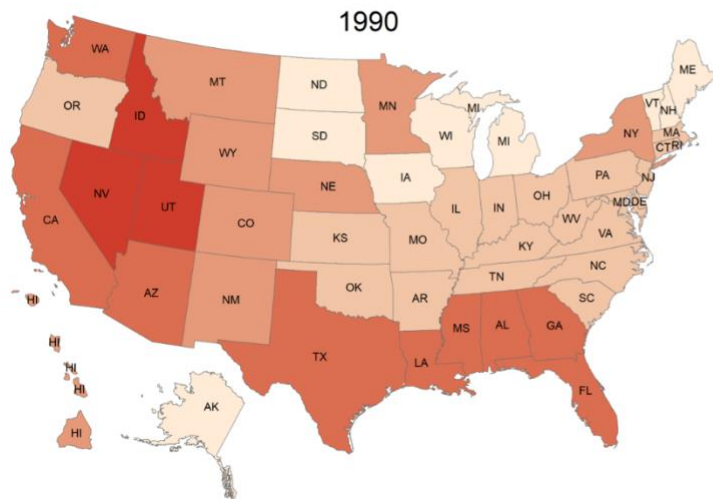


Figure 7. Clusters for Water and Weather Normalized Energy Use by State from 1985 to 2015.

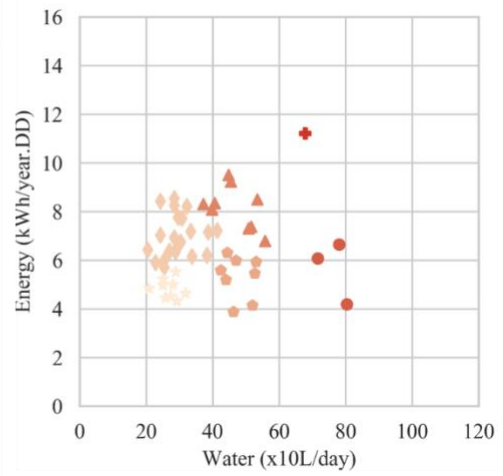


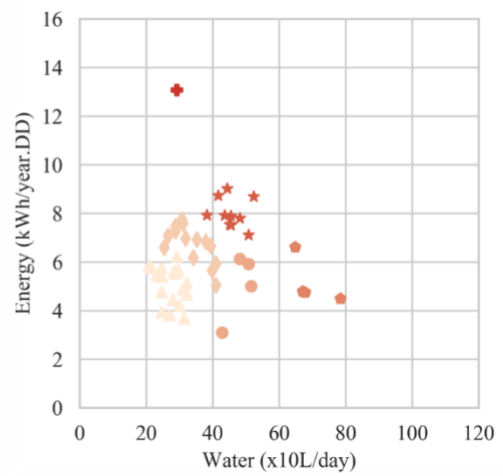
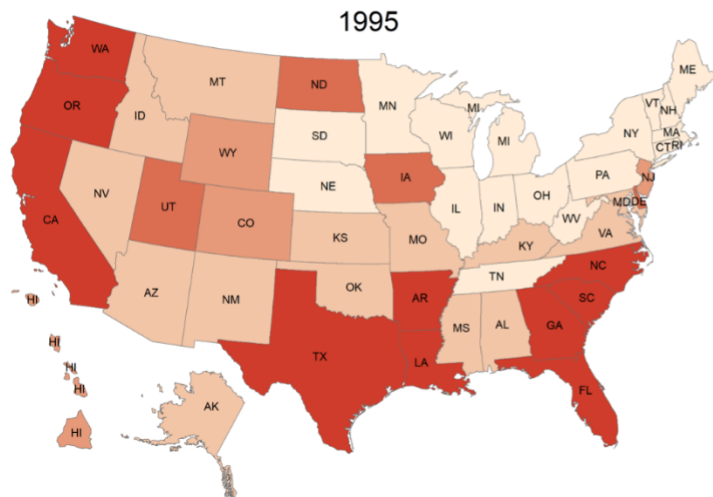
## Legend

- Low Water - Medium Energy Consumption
- Medium Water - Medium Energy Consumption
- Medium Water - High Energy
- High Water - Medium Energy
- High Water - Very High Energy Consumption



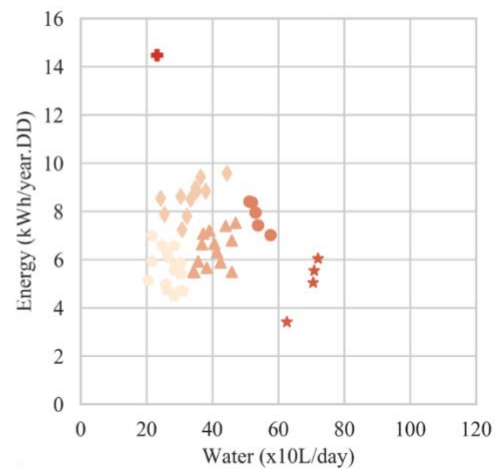
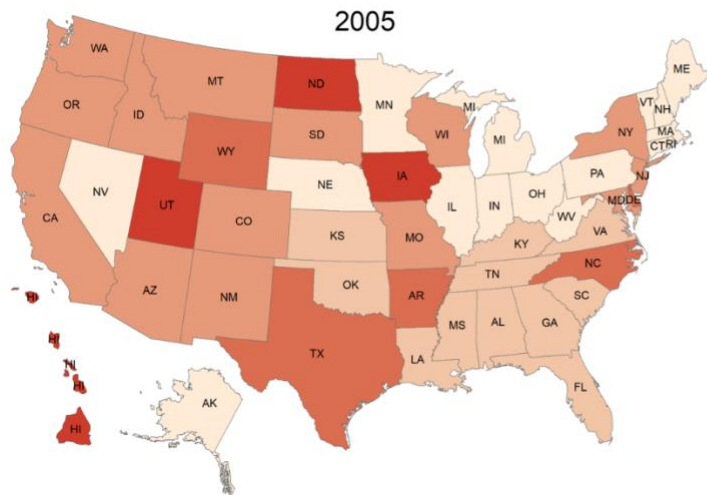
## Legend





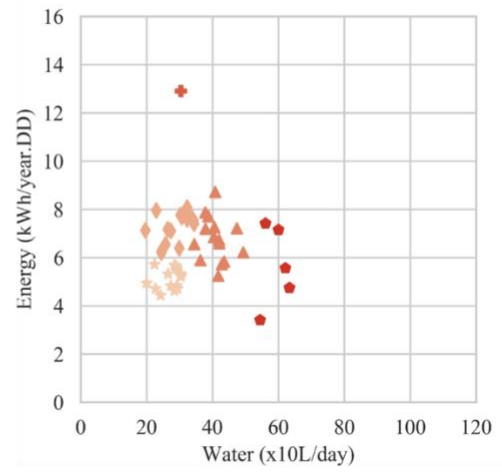
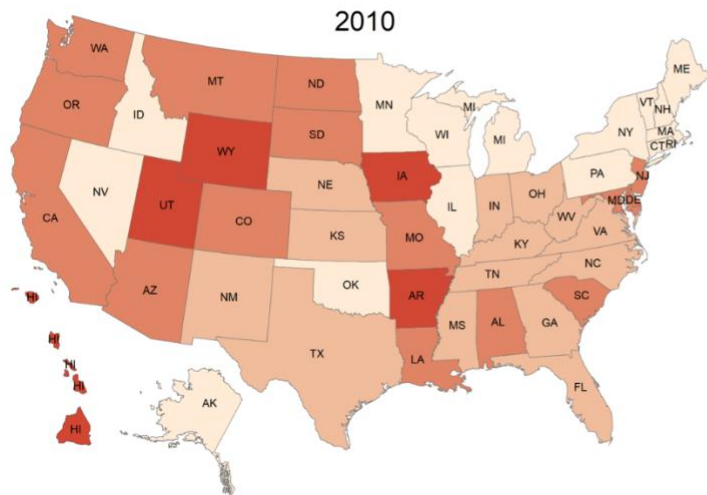
## Legend

- Low Water - Low Energy Consumption
- Low Water - Medium Energy Consumption
- Medium Water - Low Energy Consumption
- High Water - Medium Energy Consumption
- Medium Water - High Energy Consumption
- Low Water - Very High Energy Consumption



## Legend

- Low Water - Low Energy Consumption
- Low Water - Medium Energy Consumption
- Medium Water - Medium Energy Consumption
- High Water - Medium Energy Consumption
- High Water - Low Energy Consumption
- Low Water -Very High Energy Consumption



### Legend

- Low Water - Low Energy Consumption
- Low Water - Medium Energy Consumption
- Medium Water - Medium Energy
- High Water - Low Energy
- High Water - Very High Energy Consumption

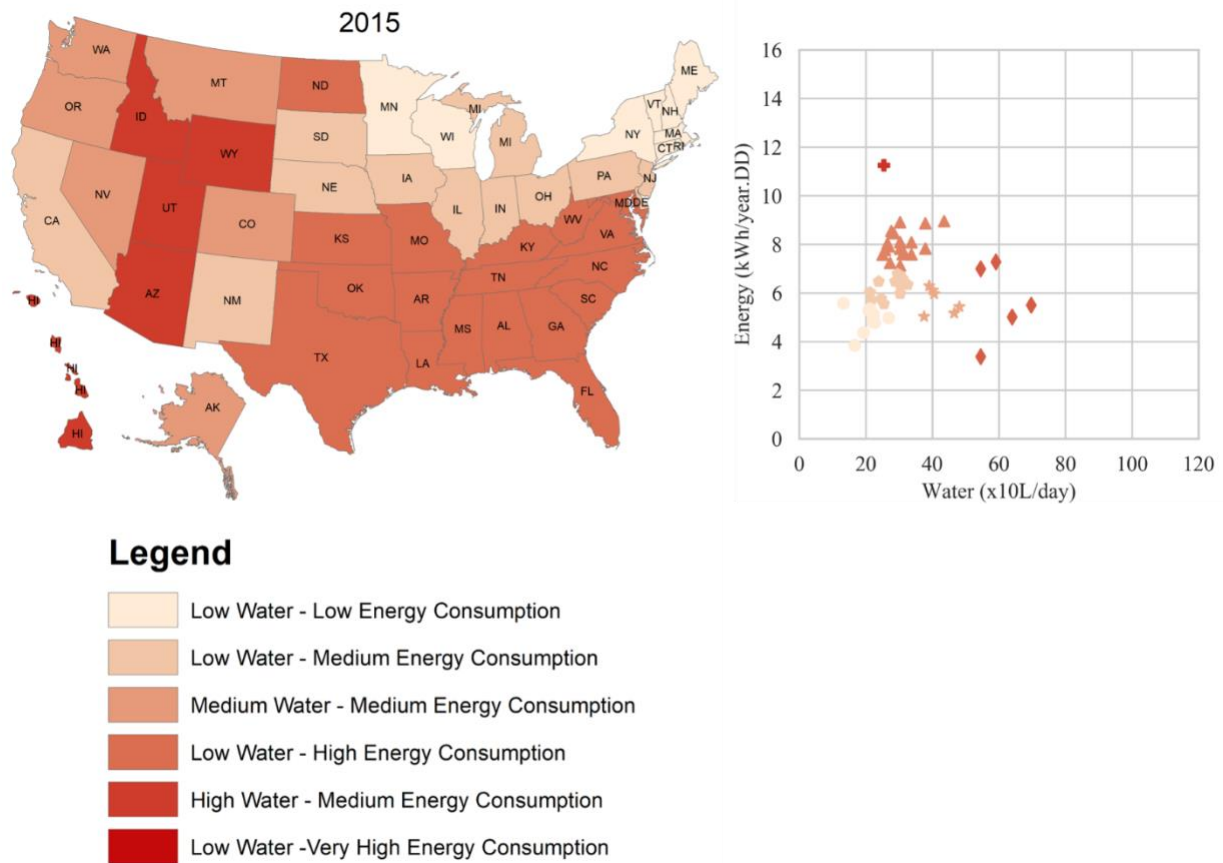


Figure 8. Clusters for Water and Weather Normalized Energy Use by State from 1985 to 2015.