Abstract: Water use patterns were explored for three small communities that are located in proximity to agricultural fields and rely on their local wells for potable water supply. High-resolution water use data, collected over a four-year period, revealed significant temporal variability. Monthly, daily, and hourly water use patterns were well described by autoregressive moving average (ARMA) models. Model development was supported by unsupervised clustering analysis via self-organizing maps (SOMs) that revealed similarities of water use patterns and confirmed the time-series water use model attributes. The inclusion of ambient temperature and rainfall as model attributes improved ARMA model performance for daily and hourly water use from $R^2 \approx 0.86–0.87$ to $0.94–0.97$ and from $R^2 \approx 0.85–0.89$ to $0.92–0.98$, respectively. Water use predictions for an entire year forward in time was feasible demonstrating ARMA models’ performance of (i) $R^2 \approx 0.90–0.94$ and average absolute relative error (AARE) of $\approx 2.9–4.9\%$ for daily water use, and (ii) $R^2 \approx 0.81–0.95$ and AARE $\approx 1.9–3.8\%$ for hourly water use. The study suggests that ARMA modeling should be useful for analysis of temporally variable water use in support of water source management, as well as assessing capacity building for small water systems including water treatment needs and wastewater handling.

Keywords: water use patterns; small communities; potable well water; autoregressive moving average (ARMA) model; self-organizing map (SOM)

1. Introduction

The freshwater sources such as rivers, lakes, reservoirs, and groundwater are increasingly being utilized worldwide [1,2]. Population growth along with increased water demand by industry, intensive agriculture, and the domestic sector are leading to excessive withdrawals from the various freshwater supplies, thus increasing water stress in various regions of the globe. Moreover, contamination of water supplies has exacerbated the situation as critical water sources are now impaired [3]. For example, in California, impaired groundwater contamination and excessive water salinity are severe in communities in agricultural regions [4,5]. Nearly 95% of the population in the communities in San Joaquin Valley, California, relies on groundwater for its drinking water needs. In this region, there are communities whose water supplies are contaminated by high nitrate levels, which is attributed, in part, to intensive agricultural activities [4,5] and impact of septic systems. In the agricultural regions, small and disadvantaged communities (i.e., with a community median household income of less than 80% of the state annual median household income), who rely on groundwater as their only potable water source, are the most severely impacted.

Small communities with impaired local well water and lack of feasible (or timely) connection to a centralized water system [6,7] can potentially opt for wellhead water
treatment as a mid- or long-term solution to providing safe drinking water [4,8]. However, sufficient data and models of high temporal resolution (hourly to seasonal variability) for forecasting small communities’ water use patterns are critical in order to establish (i) community water system design and operational specifications; (ii) water storage capacity; (iii) water treatment system treatment for upgrading community water quality as needed; (iv) handling of sanitary water; and (v) overall community planning (e.g., expansion and water system infrastructure upgrade).

There is a large body of literature on the analysis of water consumption data, at various temporal scales (i.e., monthly, weekly, and daily), and associated models for urban environments via machine learning (ML)-based models [9]. The existing studies, however, have focused primarily on describing overall city (or geographical region) water usage [9]. For example, Avni et al. [10] presented an approach to analyze average monthly water demand patterns based on classification with K-means clustering at the scale of cities, farming communities, various industries, and agrarian and communal settlements, located throughout Israel. The above study suggested that data-driven models of water use can be developed where similarities exist of water use patterns. In another study, models of total monthly urban water use have been reported based on fuzzy inference systems (FIS) that included an adaptive neuro-fuzzy inference system and a Mamdani fuzzy inference system [11]. The above water use models were with the predictive performance of $R^2$ of 0.75 for the city of Izmir [11], respectively. Models of water use were also reported for the city of Izmir based on generalized regression neural networks (GRNN), feed forward neural networks (FFNN), radial basis neural networks (RBNN), and multiple linear regression (MLR) [12]. Overall, models for monthly average urban water consumption, based on 1997–2006 data, demonstrated predictive accuracy quantified by efficiency (E) metric of 0.89 (representing the relative magnitude of the residual variance of model predictions relative to the measured data variance; the values of CORR and E close to 1.0 indicate good model performance), and normalized root mean square error (NRMSE) of ~0.07. In another study, data-driven back propagation neural network (BPNN)-based models were proposed for urban peak of weekly water demand for the town of Nicosia, Cyprus, with the population of ~200,000 [13], based on 2002–2007 data. These models built based on total city maximum daily water consumption in each week, and utilizing weekly maximum temperature and total rainfall as added model input variables from 2002 to 2006 (260 data points), demonstrated prediction accuracy for 2007 peak weekly city water consumption (52 data points) with $R^2$ ~ 0.94, and a root mean square error (RMSE) of ~0.12 million L/day relative to the average and maximum water use of 3.5 and 4 million L/day, respectively. In another study, water use in a part of the city of Ottawa (Canada), for a population of 30,000 for the period 1992–2002 [14], was analyzed via linear regression (MLR), linear neural network (LNN), and BPNN models. Models for peak daily water use in each week of the study period were developed, which considered the maximum temperature and total rainfall for the peak daily water based on training data for 1992–2001 (total of 460 data point). The highest prediction accuracy (for January–April 2002 test dataset of 18 data points) was for the BPNN model demonstrating $R^2$ ~ 0.81 and average absolute relative error (AARE) of ~0.12%.

Simpler linear regression models for daily water use (Gallons per person) in large communities, in support of establishing water management strategies, have also been developed for the Swindon Area of the Thames Water Utility, UK. (population: 0.19 million). The approach introduced simple linear regression—Partial least squares (PLS) regression econometric model—that included economic variables such as water price, household income, occupancy rate, as well as meteorological information [15]. In another study, BPNN models, along with a shuffled complex evolution metropolis (SCEM-UA) algorithm, regression, and adaptive neuro-fuzzy inference system (ANFIS), were developed to describe water demand in “Area Pilota” of Catania, Italy (population: 50,000). The model developed for daily water use averaged per person for the period of 2003–2004 [16]. Training and testing data comprised merely 200 and 65 data samples, respectively, and demonstrated
the performance of RMSE ~2.34 L/person/day relative to the average and maximum water use of 110 L/person/day and 139 L/person/day, respectively.

Predictive GRNN-based models for daily water consumption, incorporating meteorological data (i.e., average daily temperature, daily humidity, and total daily rainfall) were reported for the city of Al-Khobar (population: 455,500) in Saudi Arabia [17], based on (February 2009–October 2009) training and test data interspersed for the same period, and demonstrated the predictive performance of $R^2$ ~0.9. In another study, BPNN models of daily and hourly water usage were demonstrated for 19 different buildings from eight North American cities [18]. The above work, based on a single week of training data and subsequent testing with one week of data, demonstrated predictive performance, for single building hourly and daily water use, of AARE in 5–11% and 3–5%, respectively. In an earlier study [19], Sugeno fuzzy time series analysis [20] and autoregressive moving average (ARMA) models [21] were developed for monthly water consumption in Istanbul (12 million population), which was reported to be in the range of 10–100 million (m$^3$/year). Model training was based on a dataset spanning a period of 7 years (1995–2002). Model validation was for a period of 18 months (2003–2004) demonstrating performance of RMSE of 1.9 million (m$^3$/year) and 2.0 million (m$^3$/year) for the above two models, respectively. It is also noted that the population water use in Kuwait was reported in a study [22] that utilized a simple linear ARMA-based model in which one year-forward of water forecasting was based on the previous year consumption. Forecasting of water use was reported for the period of 2004–2025 based on water consumption data for 1954–2003. The above study also reported pair-wise correlation of water consumption and various socioeconomic factors (e.g., residence type (villa or apartment), average house size, number of household occupants, number of cars in the household, number of weekly laundry activities, weekly number of showering/bathing per household, and household monthly income). The analysis demonstrated a low level of correlation, which may suggest that water consumption may depend on multiple factors in a non-linear manner.

Relative to large urban centers, analysis and models of water use in small remote communities have been limited owing to the lack of time-series water use data. Here, we note that the estimates of household potable water use in small communities (30–400 households, 100–2400 people) for laundry, and personal hygiene have been highly approximate [23–25] given that real-time water metering data are often lacking. It is also noted that the compilation of water use data for small communities has been typically based on questionnaire and telephone surveys [23–25]. Water use data at a high temporal resolution are lacking for small communities that are not part of a centralized water distribution system.

Water use is expected to vary temporally, and thus time-series water use data are critical, particularly for communities that rely on well water, to assess needed water storage, water treatment capacity (if needed), and operational protocols. Although various ML techniques are presented in the literature, primarily for modeling water use in large urban regions, the development of robust predictive ML models is challenging when confronted with complex high-resolution time-series patterns. Additionally, models such as GRNN, ANN, BPNN, and PLS entail high complexity with a large hyperparameter space. This poses further challenges to the adaptability of such models, particularly for rapid predictions, model update, and transfer learning. Here, we note that the objective should be to arrive at a model (irrespective of the model parameter space) whereby the existing model can be used for sites of similar characteristics and where model retraining can be accomplished only based on the newly acquired data. In this regard, ARMA models have the advantage of requiring only two hyperparameters (autoregressive and moving average coefficients) [21]. Therefore, ARMA-based models can provide rapid prediction (with respect to computational time) with significantly lower training time relative to BPNNs models.

Accordingly, the current study presents a data-driven modeling approach to describe and forecast water use for small communities. The approach was explored for three small, disadvantaged communities of farm laborers and day workers located in the agricultural
region in Salinas Valley, California. Extensive multi-year high-resolution time-series water use data were compiled for each community via wireless water meters. Water use patterns were first explored at hourly, daily, and monthly resolutions via self-organizing maps (SOM) and Spearman coefficient of correlation analysis. This was followed by data-driven ARMA models considering the time of day, day of the week and month, and the daily ambient temperature and rainfall as model inputs. The models were then assessed with respect to forecasting small community water use patterns.

2. Materials and Methods

2.1. Workflow

Water use patterns in small, disadvantaged communities in the agricultural region of Salinas Valley, California were explored along with time-series models developed as per the workflow described in Figure 1. Water use data were obtained via smart water meters from three small communities, over a four-year period. The time-series water use data were initially explored by self-organizing maps (SOM) and also pair-wise correlations (Spearman coefficient). Subsequently, autoregressive moving average (ARMA) predictive models (Section S1, Supplementary Materials) were developed for water use patterns at different temporal scales. The water use patterns were analyzed to (i) assess the similarity of water use patterns among the different communities; (ii) evaluate the relevant attributes for describing water use patterns, including climate metrics, i.e., daily and monthly low/high temperature (°C) and rainfall (inches/day); and (iii) establish predictive time-series models for forecasting water use patterns.

![Figure 1. Workflow of water use data compilation, preprocessing, exploration, and model development.](image)

2.2. Study Area and Water Use Data Compilation and Preprocessing

Water use data for three small communities in Salinas Valley, California labeled as Sites A, B, and C (Table 1) were compiled over a five-year period (October 2015–October 2020). The three communities having 8–11 residential units (18–36 residents) (Table 1) are in the midst of agricultural fields. The communities rely on local wells for domestic potable water supply and their wastewater is managed in a local septic system [4].

| Study Community | Site A | Site B | Site C |
|-----------------|--------|--------|--------|
| Number of residential units | 11 | 8 | 10 |
| Population | 16 | 36 | 34 |
| Community area (m²) | 4200 | 9000 | 11,440 |
| Max water usage (gal/person/day) | 278 | 293 | 196 |
| Min water usage (gal/person/day) | 14 | 28 | 13 |
| Ave water usage (gal/person/day) | 45 | 85 | 44 |

The collection of water use data was achieved via smart (wireless) meters (Spectrum 88DL 1.5”, Metron-Farnier, Inc., Boulder, CO, USA) installed at the community main distribution line from their water delivery pressure tanks. Periodic water usage data
(i.e., volume used) was transmitted to a centralized data storage server at regular 5 min intervals. The water use dataset, which is available online (see Data Availability Statement), was utilized to determine the cumulative hourly and daily water use. Daily and hourly temperatures, as well as rainfall data for the study region, were obtained from National Oceanic and Atmospheric Administration’s climatic databases [26]. A summary of monthly rainfall and monthly average of daily low and high temperatures is provided in Table 2.

Table 2. Meteorological data for three sites’ average, from 2016 to 2020 1.

| Site | Temperature (°F) | January | February | March | April | May | June | July | August | September | October | November | December |
|------|-----------------|---------|----------|-------|-------|-----|------|------|--------|-----------|---------|----------|----------|
| Site A | Average low temperature | 41 | 43 | 45 | 46 | 50 | 53 | 55 | 55 | 54 | 50 | 43 | 40 |
|       | Average high temperature | 61 | 63 | 64 | 66 | 68 | 70 | 72 | 72 | 74 | 73 | 64 | 61 |
|       | Rainfall (inch) | 2.6 | 2.4 | 2.2 | 1.1 | 0.3 | 0.1 | 0.1 | 0.1 | 0.6 | 1.4 | 2.4 |
| Site B | Average low temperature | 36 | 37 | 38 | 41 | 45 | 50 | 54 | 55 | 52 | 46 | 39 | 36 |
|       | Average high temperature | 64 | 66 | 68 | 74 | 79 | 88 | 94 | 95 | 91 | 82 | 72 | 65 |
|       | Rainfall (inch) | 3.0 | 3.5 | 3.0 | 0.6 | 0.2 | 0.0 | 0.0 | 0.1 | 0.3 | 0.4 | 1.2 | 1.6 |
| Site C | Average low temperature | 41 | 42 | 43 | 44 | 48 | 51 | 53 | 53 | 52 | 48 | 43 | 40 |
|       | Average high temperature | 62 | 64 | 66 | 70 | 74 | 77 | 77 | 78 | 79 | 75 | 67 | 61 |
|       | Rainfall (inch) | 4.6 | 4.4 | 3.6 | 1.5 | 0.6 | 0.1 | 0.0 | 0.2 | 1.1 | 2.4 | 4.1 | 4.1 |

1 Source: (NOAA, 2020) [26].

Clustering analysis via SOM of water use was carried out based on the normalized data (i.e., to reduce data skewness) within 0–1 range as \((x_i - x_{\text{min}})/(x_{\text{max}} - x_{\text{min}})\) for \(i = 1, 2, \ldots, N\) where \(N\) is the number of data samples (Table 1). SOM clustering utilizes competitive learning that preserves the topological structure of the input space while representing the output in a lower dimension (i.e., 2-D map of cells within SOM clusters). SOMs dimensionality reduction through its discretized 2-D representation was utilized for preliminary feature selection to identify attributes of significance for ML model development. SOM was carried out for water use data for each month where the resulting 2-D SOM map (Section 3.3) represents water use per day of the week as indicated in the SOM cells. Proximities of cells in the SOM map representing the days of the week are indicative of their similarities in terms of the volume of water used. Cells are also grouped into clusters of (indicated in different colors) of similar range of daily water use. Additionally, monthly water use data were also aggregated with climate metrics (temperature (°F) and rainfall (inches)) to assess the significance and relevance of climate metrics for ARMA model development. Thus, the SOM clusters provide an indication of the relationship of climate metrics with water use as visualize by clusters with the months shown in SOM cells grouped based on daily water use for the specific month, temperature and rainfall.

Based on interviews with the residents in the study communities, it was determined that the few identified data outliers were associated with a few instances of an unusual level of car washing and community garden irrigation over brief periods of time. It is stressed, however, that data outliers were retained and included in the ARMA models (Section 2.3) given their capability to handle outliers. In addition, since the ARMA model combines auto regression and moving average as subsequent data fitting steps, variations in the reported data were robustly represented without the need for data preprocessing. Therefore, raw data without pre-normalization were directly used in the ARMA model development.

2.3. Study Area and Water Use Data Compilation and Preprocessing

Data exploration was carried out for water usage trends from hourly to monthly resolution over the course of the year. In the initial analysis, water use patterns at daily and monthly scales were evaluated based on SOM clustering to identify similarities among the communities throughout the months of the year. Water usage patterns for the study communities revealed temporal irregularity (i.e., the variance was non-stationary for the time-series data). Thus, water use data for the study sites followed non-stationary stochastic patterns with non-uniform variance. Accordingly, for ARMA model development [21,27] for water consumption, non-stationarity was removed via second order differencing [28].
The ARMA models were based on two polynomials (i.e., the first polynomial as auto-regressive and the second as moving averages). The auto-regressive (AR) polynomial constitutes the autoregressive model at a predefined order \( p \) describing the dependence of the variable (e.g., water consumption over a specified time period) on its values in a previous time. The moving average (MA) polynomial describes the linear dependence of the forecast errors resulting from the autoregressive model on the second predefined order \( q \) (Section S1, Supplementary Materials) [21,27]. The AR and MA polynomials were combined considering both the variable linear relationship and linear dependence of the forecast errors [29]. ARMA model parameter tuning was carried out based on a grid search for \( p \) and \( q \) in the range of \([10–200]\) and \([10–300]\), respectively, with incremental 2-step increase. The optimal \( p \) and \( q \) values were 66 and 72, respectively, for the hourly, and correspondingly 18 and 24 for the daily ARMA models. Model development for daily and hourly water use was based on training and test data for October 2015–December 2019 and January 2020–December 2020, respectively. Depending on the model resolution, input attributes included the hour, day, week, month of the year, in addition to the daily and hourly total rainfall, and low and high daily and hourly and ambient temperatures.

Performance of the ARMA models was quantified by R-squared and the average absolute relative error (AARE). R-squared, representing the proportion of the dependent variable (water consumption) variance predictable based on the independent variables, is given as
\[
R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2},
\]
and
\[
\text{AARE} = \left(\frac{1}{N}\right)\sum_{i=1}^{N} \left| \frac{(y_i - \hat{y}_i)}{y_i} \right|,
\]
where \( N \) is the number of data test samples, and \( y_i, \bar{y} \) and \( \hat{y}_i \) are the observed, average, and predicted values, respectively.

3. Results and Discussion
3.1. Water Use Data

Daily water consumption, averaged over each month (Figure 2) of the year, varied significantly in each community, with the highest water use typically on weekends (Saturday and Sunday). Communities B and C had higher peaks of daily water use relative to community A. Both communities B and C had home gardens, and irrigation water use may partially explain the above differences. Hourly per person water consumption, averaged annually for each weekday, (Section 2—Methods), displayed some water use similarities among the three communities but also starkly different patterns (Figure 3). As illustrated for the year 2019, the highest hourly water use was at hours 7:00, 19:00, and 12:00 for sites A, B, and C, respectively. Another seemingly high water use was at hour 20:00 and 17:00 for sites A and C, respectively. Hourly per person water similarities were also noticeable during 7:00–10:00 h and 17:00–19:00 h for site A, and 7:00–8:00 h and 17:00–20:00 h (Sunday being the exception) for site B.

Daily water use data revealed trends, indicating high water use for Saturday and Sunday (Figure 2), consistent with the highest hourly water consumption periods typically occurring during the same days (Figure 3). For site A, the highest hourly water use was for weekends (Saturday and Sunday) at around 10:00. Site B revealed a similar trend on Sunday, with the exception of Saturday in which water usage peaked at around 18:00. Similar water use peak (up to 2.5 gal/person/day) was during the early weekday mornings (4:00–7:00) for site A. Notably, increased water use for sites B and C, for all days, was during 5:00–7:00 with an immediate decline at around 9:00. For sites B and C, an increasing water consumption trend from 1 to 4 gal/person/day, was observed (Figure 3) from 9:00 to 3:00 during weekdays.
Figure 2. Daily water use (gal/person/day) for the three study communities for the year 2019. Each data point is labeled 1–7 to designate the day of the week starting with Sunday being day 1. Very High daily water use was encountered primarily on Saturday and Sunday of each week. (Characteristics of the three communities, A–C are provided in Table 1).
Figure 3. Average hourly water usage for the year 2019. The indicated hourly data for each day of the week represent averages for the given day over the entire year. High water use occurs primarily during the time period of about 7–23. As expected, there is low water use occurring during the nighttime/early morning period of 23–5.
3.2. Similarity Analysis and Exploration of Water Use Patterns

Visualization of similarities of water use patterns via SOM clustering analysis was carried out based on the accumulative water use data over each day for each month of the year (Figure 4). In SOM map the collection of contiguous cells shown in the same color represents a cluster which signifies the levels of daily water use such that the days of similar water use volume appear in adjacent cells. From the 2-D SOM representation of data similarity (represented by the proximities of cells within the map), it is seen that [Friday, Saturday, Sunday] were the days of relatively higher water usage for the majority of the months (i.e., appearing in clusters with higher color intensity (red)). However, a subtle difference is evident among certain months, as seen, for example, for February where in Cluster V, which is of the highest water use, [Sunday, Tuesday, Thursday] are the days of highest water usage. Moreover, for the highest water use Cluster II, [Saturday, Sunday, Monday] are of the highest water use days appearing at 75% frequency. In June, the highest water usage was for [Friday, Saturday] with 100% occurrence in Cluster IV. However, the highest water usage in June was for [Monday, Wednesday, Thursday] appearing (at a frequency of 100%) in Cluster VI. There are also other exceptions as observed for July where the highest water usage was Thursday (30% occurrence), while in May, the highest water usage was on Sunday (also at 50% occurrence). Interestingly, Tuesday and Wednesday typically appear in clusters of low water usage for the majority of the months. The high water use during the months of June–August shown in Table 3 (integrated with a heat map for better visualization) is also apparent in the clustering analysis shown in Figures S1–S3 (Supplementary Materials). These latter figures illustrate that [Friday, Saturday, Sunday] frequently appeared in the clusters of highest water use (red clusters) for the months [June, July, August]. SOM visualizations in Figures 4 and 5 illustrate the higher and lower cumulative water use during the months of [June, July, August] and [December, January], respectively, as also shown in Table 3 for these two sets for the period of 2015–2020. As noted in Section 3.3, there is a strong correlation of water use with daily and hourly temperature (°F) and rainfall (inch).

Table 3. Water use (gallons/day) for the three sites for the months of June–August, January and December for the period of 2015–2020.

| Site A | Site B | Site C |
|--------|--------|--------|
|        | Sunday | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday |
| January | 826.3  | 760.6  | 760.9  | 538.2     | 594.0    | 640.3  | 826.2    |
| June    | 1008.0 | 758.2  | 638.1  | 777.0     | 825.5    | 800.5  | 900.6    |
| July    | 1076.6 | 854.9  | 816.4  | 888.5     | 893.6    | 920.4  | 999.8    |
| August  | 988.1  | 773.8  | 795.6  | 767.9     | 727.9    | 975.5  | 974.6    |
| December| 664.2  | 661.2  | 557.6  | 506.7     | 514.9    | 557.1  | 590.8    |

| Site B | Site C |
|--------|--------|
|        | January | April | May | June | July | August | September | October | November | December |
|        | 1963.2  | 1563.7 | 179.7 | 1944.3 | 1914.7 | 1951.3 | 1951.8 |
|        | 4186.0  | 4346.6 | 4145.2 | 4261.0 | 4148.9 | 2768.6 | 4815.0 |
|        | 5554.6  | 4644.6 | 3304.6 | 3685.6 | 4024.2 | 5023.5 | 5940.0 |
|        | 4185.2  | 2801.3 | 2753.2 | 4715.1 | 3140.7 | 5133.0 | 5322.1 |
|        | 3008.2  | 3299.1 | 2115.6 | 2119.9 | 2325.0 | 2777.1 | 2933.5 |

Water use magnitude (from lowest to highest water use) is indicated by both the numerical data and the cell colors as per the color scale (right) where the red and green colors represent the highest and lowest water use, respectively.

1 Water use magnitude (from lowest to highest water use) is indicated by both the numerical data and the cell colors as per the color scale (right) where the red and green colors represent the highest and lowest water use, respectively.
Figure 4. SOM clustering depiction of monthly daily water patterns with respect to each month based on the overall communities’ water consumption, temperature, and rainfall dataset of the period October 2015–December 2019. In each cluster, days with highest percent of occurrence are indicated as {days} with the relative occurrences (percentage) of the day in the cluster based on the level of water consumption. Clusters are colored, on a normalized scale of 0–1, as per the color bar (right), from lowest (blue) to highest (red) water usage.
3.3. Correlation of Water Use Patterns with Meteorological Conditions

Water usage has been shown to correlate with ambient temperature and rainfall [14–16]. The monthly average of the minimum and maximum daily ambient temperatures and total rainfall have a monotonic relationship with each other (i.e., increase in temperature shows decrease in rainfall and vice versa), as shown in Figure 5 (left). Additionally, Figure 5 (right) also illustrates that daily water use per person (i.e., hourly data accumulated over each day of the month) correlates with the daily ambient temperature and rainfall for each of the three study communities. The hourly water use also correlates with ambient hourly high temperature, low temperature, and rainfall, with a Spearman correlation coefficient of 0.85–0.91, 0.83–0.86, and 0.69–0.71 for high temperature, low temperature, and rainfall, respectively (Table 4). Monthly water use increased with rising temperature and vice versa, while the converse correlation was observed with rainfall. Daily water use per person in the

Figure 5. (Left) Monthly average daily per person water use and monthly average low, $T_{\text{low}}$, and high, $T_{\text{high}}$, temperatures ($^\circ\text{F}$) and rainfall (inches) profiles in 2019 for the three study sites. (Right) SOM representation of the correlation among water usage and [temperature, rainfall] on monthly scale. Temperature and rainfall are correlated at monthly scale with [June, July, August, and September] being of higher temperature and lower rainfall, respectively. Higher water use (gallons/person/day) is correlated with higher temperature and low rainfall (red clusters at the bottom) as visualized in the clusters of high ranges (red) along with the influenced factor (lower rainfall values). The W, T, and R in the right figure refer to water use, temperature, and rainfall.
three communities was generally higher during the hotter and drier months of June through August for all the communities. Although there were differences in daily average water use per person for each month, in general, water utilization correlated with higher temperature and correspondingly lower rainfall. It is noted that higher water use in communities B (85 gal/person/day) and C (61 gal/person/day) relative to A (35 gal/person/day) was likely due to the greater use of water for garden irrigation in the two former agricultural communities (comprising of single-family homes), relative to community A being a small, isolated apartment building complex.

Table 4. The Spearman coefficients between water usage, temperature, and rainfall.

| Spearman Coefficient | Water Usage | High Temperature | Low Temperature | Rainfall |
|----------------------|-------------|------------------|-----------------|----------|
| Daily                | Water usage in Site A | 0.86             | 0.81            | −0.72    |
|                      | Water usage in Site B | 0.88             | 0.84            | −0.79    |
|                      | Water usage in Site C | 0.82             | 0.78            | −0.75    |
| Hourly              | Water usage in Site A | 0.89             | 0.83            | −0.69    |
|                      | Water usage in Site B | 0.91             | 0.86            | −0.71    |
|                      | Water usage in Site C | 0.85             | 0.84            | −0.70    |

1 Temperature and rainfall data are as per Table 2.

The correlation of water usage patterns with the low and high daily temperatures and rainfall events were also quantitatively assessed via the Spearman coefficient [30], as shown in Table 4. The Spearman coefficient was chosen as the metric for attribute correlation due to the non-linearity and monotonicity (i.e., the relationship between temperature and rainfall and their impact on the water use). As indicated in Table 4, a strong positive correlation of water use with temperature (i.e., higher use at higher temperature) was determined (Spearman coefficient of 0.78–0.88) for the three communities. Water use had a strong but negative correlation (Spearman coefficient in the range of −0.72 to −0.79) with rainfall (i.e., decreased water use with higher rainfall).

Exploratory analysis of the water use data (via SOM and Spearmin coefficient analyses) suggested that water use correlates with ambient temperature and rainfall, consistent with other studies on water use in various regions [31]. Accordingly, these meteorological attributes were included in the developed ARMA models (Section 3.4).

3.4. Data-Driven ARMA Models for Water Use Patterns

The ARMA models for daily and hourly water use were developed for the three communities based on data compiled over a period of four years, following the workflow presented in Figure 1. As an example, model performance for daily and hourly water use, with and without the inclusion of meteorological parameters, is illustrated in Figures 6 and 7, respectively, for 2017 training data traces. Model performance for the entire training dataset (i.e., October 2015–December 2020) is provided in Table 5, and model validation with the test dataset (January 2020–December 2020) is presented in Section 3.5.

ARMA model performance for the complete training dataset for daily and hourly water use for the three sites was in the range of $R^2 \approx 0.94-0.97$ and 0.92–0.98, respectively, but was correspondingly significantly lower (0.87–0.89, and 0.86–0.92) when daily temperature and rainfall were excluded as input parameters. The AARE levels for the daily and hourly ARMA models were in the range of ~2.86–5.20% and ~2.55–3.88%, respectively. However, ARMA models without meteorological parameters as input attributes resulted in higher AARE 3.88–7.89% and 2.86–5.2% for daily and hourly water use, respectively. Variations of per person water use in communities A, B, and C varied by up to factors of 2–4 within each month of the year. Hourly per person water use variability was much greater ranging from no use to as high as 219, 1235, and 499 gallons/hr for Sites A, B, and C, respectively. Finally, the observed and model predicted community minimum, maximum, and average daily water use for each of the study years, as shown in Table S1 (Section S3, Supplementary Materials), are in excellent agreement.
Figure 6. Illustration of model performance for daily water use (gal/person/day) for a 2019 trace of the training dataset with (red) and without (blue) meteorological inputs vs. measured water usage for the three study sites. (Met—meteorological; W/O—without). For sites {A–C}, ARMA model training was based on water use data for the periods of October 2015–June 2019, and October 2015–December 2019, respectively.
Figure 7. Illustration of the performance of ARMA models for hourly water use (gal/person/day) in Sites A and B, and C for portion of a training dataset trace of the first week in March 2017 (black), with (red) and without (blue) meteorological input for the three study sites. (Met—meteorological; W/O—without).
Table 5. Training accuracy of ARMA models in three sites.

| Indicator | Site A | Site B | Site C |
|-----------|--------|--------|--------|
|           | No     | Yes    | No     | Yes    | No     | Yes    |
| Daily     | AARE (%) | 4.28 | 3.20 | 3.88 | 2.95 | 7.89 | 3.05 |
|           | $R^2$  | 0.87 | 0.94 | 0.89 | 0.97 | 0.86 | 0.96 |
| Hourly    | AARE (%) | 3.76 | 2.91 | 2.86 | 2.55 | 5.20 | 3.38 |
|           | $R^2$  | 0.89 | 0.96 | 0.92 | 0.98 | 0.85 | 0.92 |

Table 6. Performance of ARMA models for the test data for the three study sites.

| Performance Indicator | Site A | Site B | Site C |
|-----------------------|--------|--------|--------|
|                       | No     | Yes    | No     | Yes    | No     | Yes    |
| Daily                 | AARE (%) | 6.83 | 3.92 | 5.82 | 2.90 | 9.89 | 4.95 |
|                       | $R^2$  | 0.79 | 0.91 | 0.82 | 0.94 | 0.77 | 0.90 |
| Hourly                | AARE (%) | 7.86 | 2.95 | 6.76 | 1.91 | 9.73 | 3.83 |
|                       | $R^2$  | 0.75 | 0.93 | 0.84 | 0.95 | 0.74 | 0.91 |

3.5. Validation of the ARMA Model of Water Use Patterns

SOM analysis of water use patterns (Section 3.3) demonstrated that the study communities exhibited similar water use trends, thus suggesting that ARMA model validation could be carried out with future data (i.e., forward in time relative to the training data). Accordingly, the time series dataset for the period of the year 2020, which was not utilized for model training, was utilized for model validation, as shown in Figure 8. In the absence of temperature and rainfall as model attributes, model predictions for daily and hourly water use for the year 2020, for the three communities, was with $R^2$ in the range of 0.90–0.95. Here, we note that true water use forecasting would require meteorological data, which would essentially require a predictive model. As expected, the inclusion of temperature and rainfall as model inputs improved predictive performance in terms of $R^2$ by as high as 12% (from $R^2$ of 0.82 to 0.94) and 17% (from $R^2$ of 0.74 to 0.91) for average daily water use for Site B, and hourly data for Site C, respectively (Table 6).

The ARMA models of daily water usage (Figure 8) for sites A–C also demonstrated excellent performance (i.e., mean $R^2$ for three sites ~0.92) with climate metrics included in model inputs. ARMA model performance with the meteorological parameters included was 12% higher relative to the models without these climate parameters ($R^2$ ~0.8). ARMA models of hourly water use with climate parameters included also demonstrated good performance of mean $R^2$ ~0.93 relative to which was 13% higher when these parameters were not considered (Figure 9).
Figure 8. Illustration of validation of daily water use (gal/person/day) ARMA models with (red) and without (blue) meteorological (“Met”) for the three study sites. Note: training data for sites {A–C} were for the periods October 2015–June 2019 and October 2015–December 2019, respectively (Figure 6).
Figure 9. Illustration of validation of the hourly water usage model for a data trace during the first week of March 2020 (note: the indicated $R^2$ values are for the shown data trace). The ARMA model accounts for the month, week and day, hourly temperature, and total hourly rainfall input. Note: Training for hourly ARMA models for Sites A–C data were for the period of October 2015–December 2019 (Figure 7).
The ARMA models accompanied by similarity analysis should be particularly useful for guiding the design and management of small community water systems including wellhead water treatment (if required). The present work also suggests that ARMA models developed in the present work should prove useful for providing estimates of water use in similar communities. Such models, as a starting base, could also be refined via incremental learning as new data become available. Selection of lightweight ARMA models with tuned hyperparameters can accelerate the learning process of water use patterns for newer sites of similar characteristics via a pretrained model (transfer learning). In this regard, only newly acquired data would then be needed for updating the ARMA model via the transfer learning approach whereby model hyperparameters are reused for training the existing ARMA model for the new sites. It is noted that such an approach of transfer learning can be particularly useful when the water use dataset for the target site is limited.

4. Conclusions

Water use patterns in multiple small communities, located in Salinas Valley, California (United States), were collected over a four-year period and analyzed to assess and quantitatively describe water use patterns. Self-organizing map (SOM) clustering was used for visual depiction of similarities in water use patterns among the days of the week and months of the year. SOM data exploration of the individual sites collectively showed that [Friday, Saturday, Sunday] are days with the highest water usage. SOM analysis further demonstrated that during the week, [Tuesday, Wednesday] are typically the days of lowest water usage. Among the three study communities, the daily peak water usage was during the periods of about 7:00–9:00 and 18:00–22:00. The highest daily water use during the week was for Saturday and Sunday and highest monthly water use was during the months of June, August, and September. Given that water use represents time-series data, predictive ARMA models were developed for different time scales, for each of the study sites, based on water use training data for the period of October 2015–October 2019 and test data for the period of January 2020–December 2020. The models included input regarding population density, categorical information (hour of the day, day of the week, and associated month) and climate metrics (temperature and rainfall). The performance of the ARMA models (for each community) for daily and hourly water use, based on a year of data forward in time relative to the training data, was with $R^2$ in the range of 0.91–0.94 and 0.91–0.95, respectively, and corresponding absolute average error (AARE) of 2.9–4.95% and 1.91–3.83%. The present study suggests that there is merit in considering the ARMA type models for supporting water source management, and the design and deployment of local water systems, including the needed capacity for water treatment and wastewater handling. As suggested by the present similarity analysis of water use patterns for the three small study communities, it may be feasible to invoke transfer learning for the ARMA models to accelerate model training for similar sites, particularly when water use data may be limited. Admittedly, the development of water use models that are of a more general applicability would require specific continuous and categorical model parameters that are expanded to include, for example, details of community descriptors such as personal income, occupation, average residents per household, size of residential units and their number per community, as well as the specific source water (i.e., local well or centralized source).

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/w13162312/s1, Section S1: ARMA Modeling, Section S2: 2. SOM Analysis for water usage pattern in three communities. Figure S1: SOM depiction of monthly daily water patterns with respect to each month based on the water consumption dataset of the period October 2015–December 2019 for Site A. Figure S2: SOM depiction of monthly daily water patterns with respect to each month based on the water consumption dataset of the period October 2015–December 2019 for Site B. Figure S3: SOM depiction of monthly daily water patterns with respect to each month based on the water consumption dataset of the period October 2015–December 2019 for Site C. Additionally, the detailed water use data are available online as indicated in the Data Availability Statement.
Author Contributions: Y.Z.: Data analysis, model development, writing—draft preparation. B.M.K.: Workflow development, model review, and writing: draft preparation and review. J.Y.C.: Installation of water meters, monitoring of online data acquisition and management system, and data compilation. Y.C.: Study conceptualization, project supervision, modeling workflow review, and writing: draft preparation and review. All authors have read and agreed to the published version of the manuscript.

Funding: Data collection was funded by the State of California Water Resources Control Board, Agreement 14-255-550 (C/A 367) and the UCLA Water Technology Research (WaTeR) Center.

Data Availability Statement: Water use data in Excel file format are available in the submitted Supplementary Materials, which is accessible from DRYAD: https://doi.org/10.5068/D15D61, accessed on 17 July 2021.

Acknowledgments: This work was supported, in part, by the State of California Water Resources Control Board, Agreement 14-255-550 (C/A 367) and UCLA Water Technology Research (WaTeR) Center. The support by China Scholarship Council to Yang Zhou is also acknowledged.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Winton, D.J.; Anderson, L.G.; Rocifé, S.; Loiselle, S. Macroplastic pollution in freshwater environments: Focusing public and policy action. Sci. Total Environ. 2020, 704, 135242. [CrossRef]
2. McMahon, P.B.; Bohlke, J.K.; Dahm, K.G.; Parkhurst, D.L.; Anning, D.W.; Stanton, J.S. Chemical Considerations for an Updated National Assessment of Brackish Groundwater Resources. Groundwater 2016, 54, 464–475. [CrossRef]
3. Gunnaarsdottir, M.J.; Gardarsson, S.M.; Figueras, M.J.; Puigdomènech, C.; Juárez, R.; Saucedo, G.; Arnedo, M.J.; Santos, R.; Monteiro, S.; Avery, L.M.; et al. Water safety plan enhancements with improved drinking water quality detection techniques. Sci. Total Environ. 2020, 698, 134185. [CrossRef]
4. Choi, J.Y.; Lee, T.; Aleidan, A.B.; Rahardianto, A.; Glickfeld, M.; Kennedy, M.E.; Chen, Y.A.; Haase, P.; Chenc, C.; Cohen, Y. On the feasibility of small communities wellhead RO treatment for nitrate removal and salinity reduction. J. Environ. Manag. 2019, 250, 109487. [CrossRef] [PubMed]
5. Balazs, C.; Morello-Frosch, R.; Hubbard, A.; Ray, I. Social Disparities in Nitrate-Contaminated Drinking Water in California’s San Joaquin Valley. Environ. Health Perspect. 2011, 119, 1272–1278. [CrossRef]
6. Hargrove, W.L.; Holguin, N.; Tippin, C.L.; Heyman, J.H. The soft path to water: A conservation-based approach to improved water access and sanitation for rural communities. J. Soil Water Conserv. 2020, 75, 38–44. [CrossRef]
7. Schoeman, J.J. Nitrate-nitrogen removal with small-scale reverse osmosis, electrodialysis and ion-exchange units in rural areas. Water SA 2009, 35, 721–728. [CrossRef]
8. Brady, P.V.; Kottenstette, R.J.; Thomas, M.M.; Hightower, M.M. Inland Desalination: Challenges and Research Needs. J. Contemp. Water Res. Educ. 2005, 132, 46–51. [CrossRef]
9. Donkor, E.A.; Mazzuchi, T.A.; Soyer, R.; Roberson, J.A. Urban Water Demand Forecasting: A Review of Methods and Models. J. Water Resour. Plan. Manag. 2014, 140, 146–159. [CrossRef]
10. Avni, N.; Fishbain, B.; Shamir, U. Water consumption patterns as a basis for water demand modeling. Water Resour. Res. 2016, 51, 8165–8181. [CrossRef]
11. Firtat, M.; Yurdusev, M.A. Comparative analysis of fuzzy inference systems for water consumption time series prediction. J. Hydrol. 2009, 374, 235–241. [CrossRef]
12. Firtat, M.; Yurdusev, M.A.; Turan, M.E. Evaluation of Artificial Neural Network Techniques for Municipal Water Consumption Modeling. Water Resour. Manag. 2009, 23, 617–632. [CrossRef]
13. Adamowski, J.; Karapataki, C. Comparison of Multivariate Regression and Artificial Neural Networks for Peak Urban Water-Demand Forecasting: Evaluation of Different ANN Learning Algorithms. J. Hydrol. Eng. 2010, 5, 729–743. [CrossRef]
14. Bougadis, J.; Diduch, R. Short-term municipal water demand forecasting. Hydrol. Process. 2005, 19, 137–148. [CrossRef]
15. Froukh, M.L. Decision-Support System for Domestic Water Demand Forecasting and Management. Water Resour. Manag. 2001, 15, 363–382. [CrossRef]
16. Cutore, P.; Campisano, A.; Kapelan, Z.; Savic, D. Probabilistic prediction of urban water consumption using the SCEM-UA algorithm. Urban Water J. 2008, 5, 125–132. [CrossRef]
17. Al-Zahrani, M.A.; Abo-Monasar, A. Urban Residential Water Demand Prediction Based on Artificial Neural Networks and Time Series Models. Water Resour. Manag. 2015, 29, 3651–3662. [CrossRef]
18. Jentgen, L.; Kidder, H.; Conrad, R.H. Energy management strategies use short-term water consumption forecasting to minimize cost of pumping operations. J. Am. Water Work. Assoc. 2007, 99, 86–94. [CrossRef]
19. Altunkaynak, A.; Ozger, M.; Cakmakci, M. Water consumption prediction of Istanbul City by using fuzzy logic approach. Water Resour. Manag. 2005, 19, 641–654. [CrossRef]
20. Fiordaliso, A. A constrained Takagi-Sugeno fuzzy system that allows for better interpretation and analysis. *Fuzzy Sets Syst.* **2001**, *118*, 307–318. [CrossRef]

21. Ljung, G.M.; Box, G.E.P. The likelihood function of stationary autoregressive-moving average models. *Biometrika* **1979**, *66*, 265–270. [CrossRef]

22. Alhumoud, J.M. Freshwater consumption in Kuwait: Analysis and forecasting. *J. Water Supply Res. Technol.-Aqua.* **2008**, *57*, 279–288. [CrossRef]

23. Luo, Y.; Guo, W.; Ngo, H.H.; Nghie, L.D.; Hai, F.I.; Zhang, J.; Liang, S.; Wang, X.C. A review on the occurrence of micropollutants in the aquatic environment and their fate and removal during wastewater treatment. *Sci. Total Environ.* **2014**, *473*, 619–641. [CrossRef]

24. Baz-Lomba, J.A.; Salvatore, S.; Gracia-Lor, E.; Bade, R. Comparison of pharmaceutical, illicit drug, alcohol, nicotine and caffeine levels in wastewater with sale, seizure and consumption data for 8 European cities. *BMC Public Health* **2016**, *16*, 1035–1045. [CrossRef] [PubMed]

25. Hadjer, K.; Klein, T.; Schopp, M. Water consumption embedded in its social context, north-western Benin. *Phys. Chem. Earth* **2005**, *30*, 357–364. [CrossRef]

26. National Oceanic and Atmospheric Administration (NOAA). National Weather Service, National Centers Environmental Information. 2020. Available online: https://gis.ncdc.noaa.gov/maps/ncei/summaries/daily (accessed on 8 August 2021).

27. Li, Z.Y.; Yan, H.; Zhang, C.; Tsung, F. Long-Short Term Spatiotemporal Tensor Prediction for Passenger Flow Profile. *IEEE Robot. Autom. Lett.* **2020**, *5*, 5010–5017. [CrossRef]

28. Chen, W.; Wang, S. A 2nd-order ADI finite difference method for a 2D fractional Black–Scholes equation governing European two asset option pricing. *Math. Comput. Simul. (MATCOM)* **2020**, *171*, 279–293. [CrossRef]

29. Wang, Y.; Wang, D.; Tang, Y. Clustered Hybrid Wind Power Prediction Model Based on ARMA, PSO-SVM, and Clustering Methods. *IEEE Access* **2020**, *8*, 17071–17079. [CrossRef]

30. Joe, H. Families of m-Variate distributions with given margins and m(m−1)/2 bivariate dependence parameters. *Distrib. Fixed Marg. Relat. Top. Lect. Notesmonograph* **1996**, *28*, 120–141. [CrossRef]

31. Min, W.; Zhou, W.; Qi, T.; Li, H. Deep Supervised Quantization by Self-Organizing Map. *ACM Trans. Multimed. Comput. Commun. Appl.* **2019**, *15*, 81. [CrossRef]