Machine Learning Modeling of Water Use Patterns in Small Disadvantaged Communities: Supplementary Materials

Yang Zhou [1,3], Bilal Muhammad Khan [2,4], Jin Yong Choi [1,4] and Yoram Cohen [1,4,*]

[1] Chemical and Biomolecular Engineering Department, University of California, Los Angeles; Los Angeles, California 90095, USA

[2] Department of Computer Science and Engineering, California State University San Bernardino, 5500 University Parkway, San Bernardino, California 92407 USA

[3] Department of Automation, Shanghai University, Shanghai 200444, China.

[4] Institute of the Environment and Sustainability, University of California Los Angeles; Los Angeles, California 90095, USA

* Correspondence: profyc@gmail.com; Tel.: +1-(310)-713-1543

S1. ARMA modeling

The ARMA models provide an approach that suitable for describing time-series data [1] that portray stationary stochastic processes, relying on the combination of auto-regressive and moving average polynomials [2]. The autoregressive polynomial sub-model is at a predefined order p describing the relationship of the variable captured by a p order polynomial (e.g., cumulative daily/hourly water consumption over a specified time period in the present study) based on its previous time values. The moving average polynomial is a sub-model that describes the dependence of the forecast errors resulting from the autoregressive (polynomial) model on the second predefined order q. The ARMA polynomial models are described by linear stochastic models expressed as [2],

$$\tilde{Y}_t = \mu + \varepsilon_t - \phi_1\varepsilon_{t-1} - \phi_2\varepsilon_{t-2} - \cdots - -\phi_q\varepsilon_{t-q_t} - \cdots \quad \text{(S1)}$$

where μ is the mean of a stationary process, ε_t is an uncorrelated random variable with zero mean and constant variance, and ϕ_t, t=1,2,…, are coefficients which satisfy the following inequality,

$$\sum_{i=0}^{\infty}\phi_i^2 < \infty \quad \text{(S2)}$$

It is convenient to express Eq. (A1) in terms of a finite number of autoregressive (AR) and/or moving average (MA) components. Accordingly, with Y_t defined as $\tilde{Y}_t - \mu$, the deviation of the process from some origin, or from its mean, the AR process with order p can be expressed as follows:

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t \quad \text{(S3)}$$

and likewise, an MA process with order q can be expressed as:

$$Y_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q_t} \quad \text{(S4)}$$

Following the above, the general expression for a combined ARMA(p, q) process is defined as

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q_t} \quad \text{(S5)}$$

where φ_t and ϑ_t are coefficients that satisfy stationarity and invert ability conditions, respectively [3].

References

1. Rout, M., Majhi, B., Majhi, R. and Panda, G. Forecasting of currency exchange rates using an adaptive ARMA model with differential evolution based training. *Journal of King Saud University Computer & Information Sciences*, **2014**, 26, 7-18., [https://doi.org/10.1016/j.jksuci.2013.01.002]
2. Box, G.E.P., Jenkins, G.M., Reinsel, G.C. Time Series Analysis: Forecasting and Control, Prentice-Hall, Englewood Cliffs, *New Jersey*. 1994. [http://www.wiley.com/go/fo/permission]
3. 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. [https://doi.org/10.1109/ACCESS.2020.2968390]

S2. SOM Analysis for water usage pattern in three communities

SOM water use analysis for the individual community sites were consistent with analysis of the collective water use data for the study sites (manuscript **Figure 4**). It is noted that in the SOM analysis, while each day has a maximum of four appearances in the SOM analysis of the average daily data with respect to each month, water usage for a given day of the week in different weeks of the month may not necessarily correlate. Therefore, a day on the topological SOM structure may appear in different cells, or even different clusters. For example, Sunday for the month of April for Site A (Figure S1) appears in both the clusters of the lowest (blue) as well as highest threshold (red). Additionally, as illustrated in the manuscript (**Figure 4**), Tuesday is identified clusters of the highest (red) as well as lowest (blue) water use thresholds. SOM visualization shows that although {Friday, Saturday, Sunday} are days of higher water use, subtle differences exist in the water use data for the different sites. Overall, {Friday, Saturday, Sunday} were the days of highest water usage for 7-9 months of the year for the three sites (**Figures S1-S3**). However, there are subtle difference in water use patterns among the different months of the years. For example, in Site A for the month of May, {Tuesday} was the highest water use day (at 100% occurrence) appearing in Cluster V, followed by Cluster I in which the {Friday, Saturday} were are identified at an occurrence of 62.5%. In site B, the higher water usage is for {Friday, Saturday} in 8 months (i.e. in January, March, April, May, July, August, October and November). It is noted that irrespective of the data gaps for site C (i.e., missing data for 16 out of 48 months), the higher water usage was for {Friday, Saturday} for the 6 out of 9 for the 8 months. Interestingly,{Tuesday, Wednesday} were days of low water usage for 77-9 of the months for all three sites.

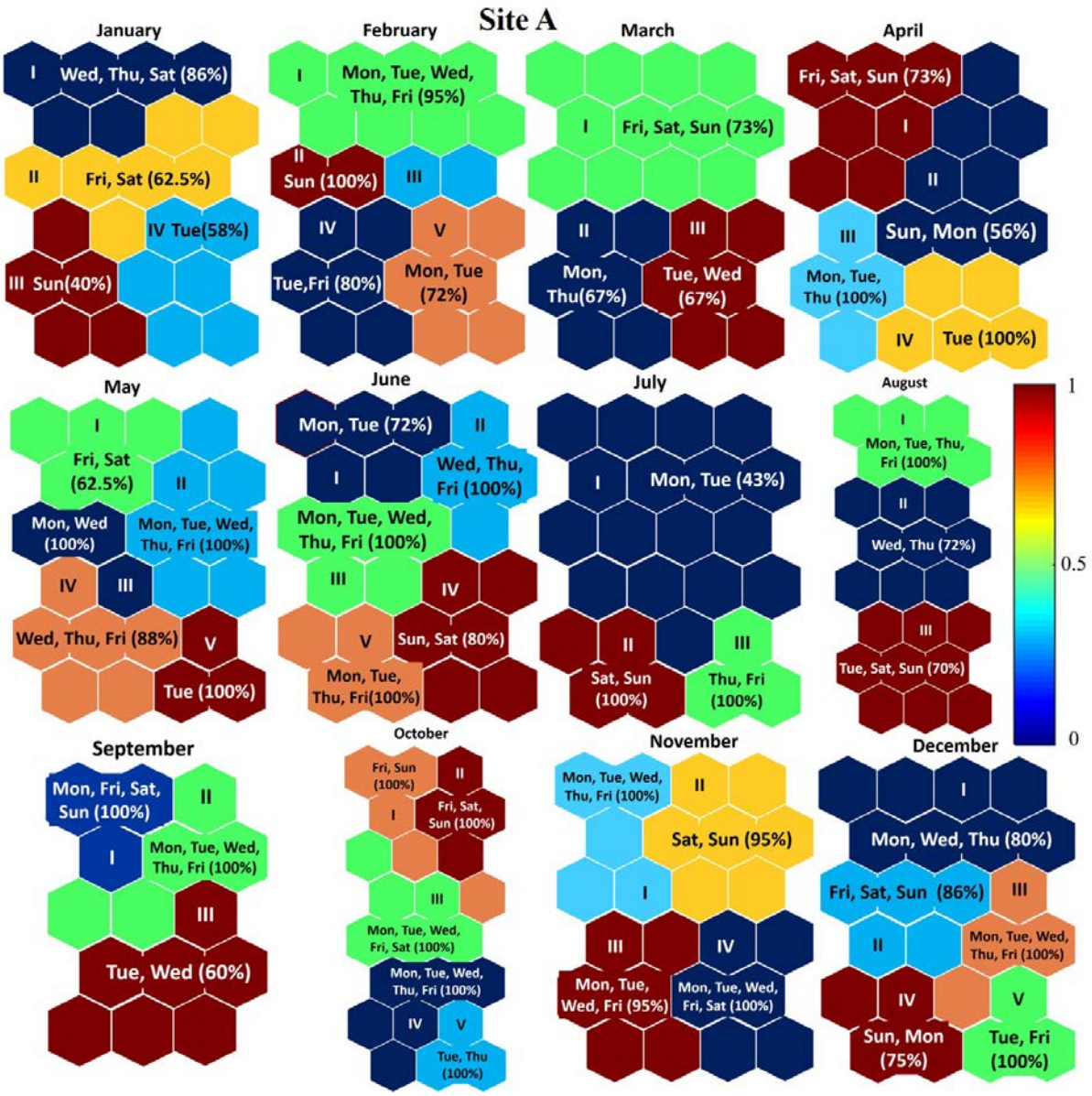

Figure S1. *SOM depiction of monthly daily water patterns with respect to each month based on the water consumption dataset of the period October 2015 - December 2020 for Site A. In each cluster, days with highest percentages are provided as {days} (relative occurrences (percentage) of the day in the cluster based on the water consumption). Clusters are colored on a normalized scale of 0 – 1 as per the color bar, where the range from 0 to 1 is given on the scale from lowest (blue) to highest (red) water usage.*

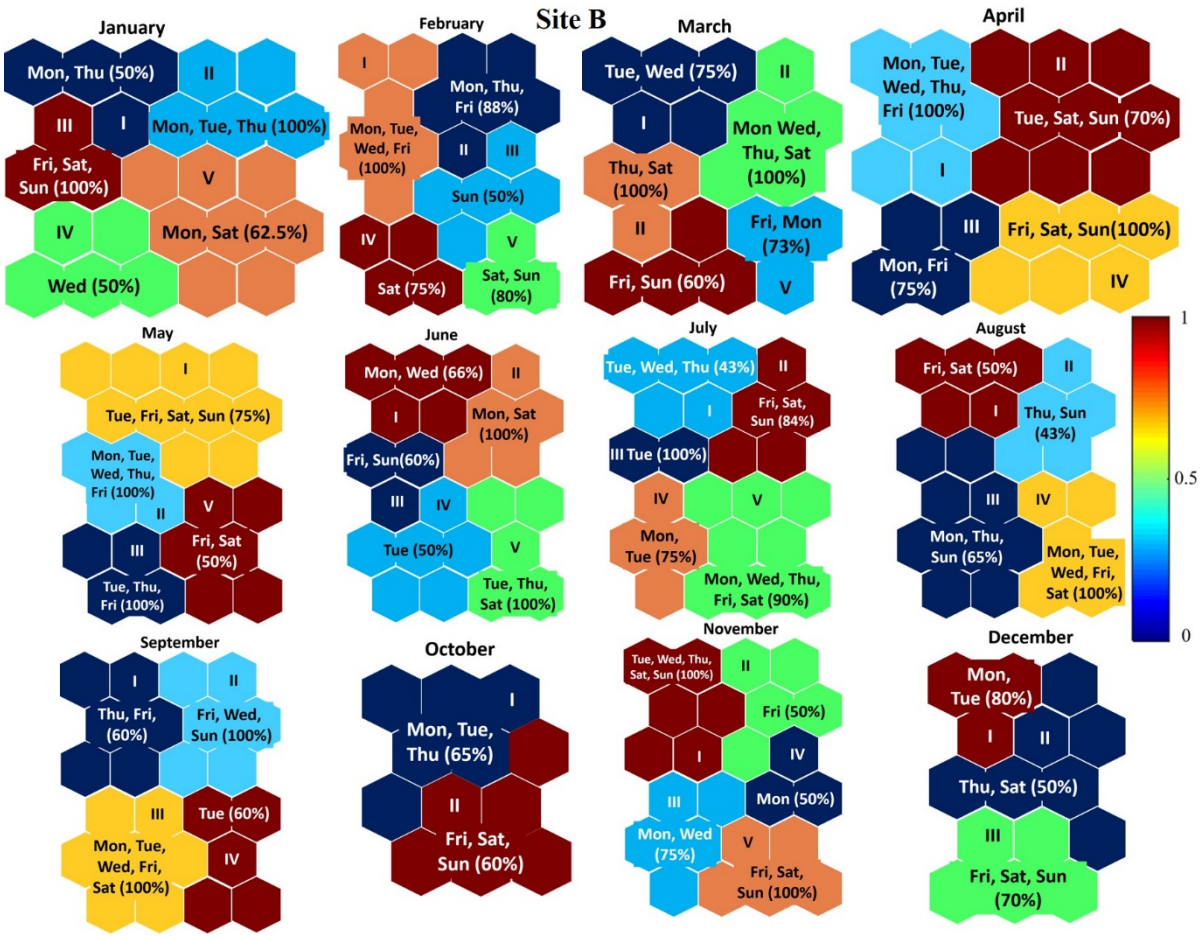

Figure S2. *SOM depiction of monthly daily water patterns with respect to each month based on the water consumption dataset of the period October 2015 - December 2020 for Site B. In each cluster, days with highest percentages are provided as {days} (relative occurrences (percentage) of the day in the cluster based on the water consumption). Clusters are colored on a normalized scale of 0 – 1 as per the color bar, where the range from 0 to 1 is given on the scale from lowest (blue) to highest (red) water usage.*

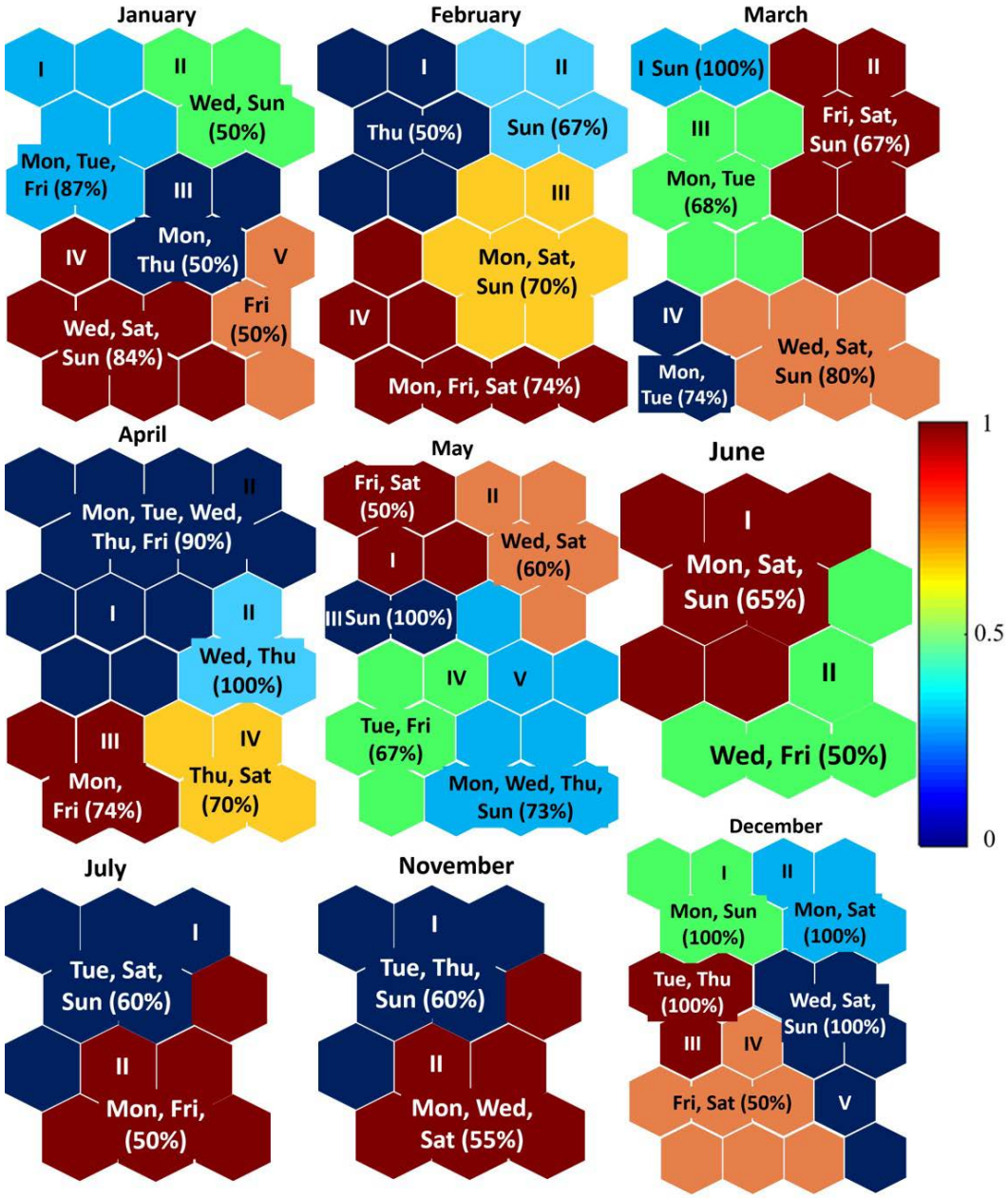

Figure S3. *SOM depiction of monthly daily water patterns with respect to each month based on the water consumption dataset of the period October 2015 - December 2020 for Site C. In each cluster, days with highest percentages are provided as {days} (relative occurrences (percentage) of the day in the cluster based on the water consumption). Clusters are colored on a normalized scale of 0 – 1 as per the color bar, where the range from 0 to 1 is given on the scale from lowest (blue) to highest (red) water usage.*

S3. Summary of community daily water consumption

Table S1. The range of observed and predicted average daily community water use for the period 2016 to 2020.

		Site A				
Year		2016	2017	2018	2019	2020
Observed (gal/day)	Max	4448	2095	860	1702	1106
	Min	287	379	339	274	271
	Ave	1118	714	585	604	568
Predicted (gal/day)	Max	4435	2085	867	1694	1113
	Min	282	372	345	224	272
	Ave	1114	710	591	600	576
		Site B				
Year		2016	2017	2018	2019	2020
Observed (gal/day)	Max	10548	8029	9561	7844	6883
	Min	1226	1255	1008	1197	1011
	Ave	3355	3144	2743	2811	2777
Predicted (gal/day)	Max	10531	8022	9555	7841	6877
	Min	1217	1251	1011	1196	1006
	Ave	3350	3139	2738	2807	2770
		Site C				
Year		2016	2017	2018	2019	2020
Observed (gal/day)	Max	2044	3220	3214	6664	5004
	Min	511	696	631	442	875
	Ave	975	1323	1469	1878	1858
Predicted (gal/day)	Max	2039	3215	3211	6627	4991
	Min	514	692	625	449	863
	Ave	975	1323	1462	1872	1851