Application of Machine Learning-based Energy Use Forecasting for Inter-basin Water Transfer Project

Sooyeon Yi[1] · **G. Mathias Kondolf**[1] · **Samuel Sandoval-Solis**[2] · **Larry Dale**[3]

Abstract

Energy use forecasting is crucial in balancing the electricity supply and demand to reduce the uncertainty inherent in the inter-basin water transfer project. Energy use prediction supports the reliable water-energy supply and encourages cost-effective operation by improving generation scheduling. The objectives are to develop subsequent monthly energy use predictive models for the Mokelumne River Aqueduct in California, US. Partial objectives are to (a) compare the model performance of a baseline model (multiple linear regression (MLR)) to three machine learning-based models (random forest (RF), deep neural network (DNN), support vector regression (SVR)), (b) compare the model performance of the whole system to three subsystems (conveyance, treatment, distribution), and (c) conduct sensitivity analysis. We simulate a total of 64 cases (4 algorithms (MLR, RF, DNN, SVR) x 4 systems (whole, conveyance, treatment, distribution) x 4 scenarios (different combinations of independent variables). We concluded that the three machine learning algorithms showed better model performance than the baseline model as they reflected non-linear energy use characteristics for water transfer systems. Among the three machine learning algorithms, DNN models yielded higher model performance than RF and SVR models. Subsystems performed better than the whole system as the models more closely reflected the unique energy use characteristics of the subsystems. The best case was having water supply (t), water supply (t-1), precipitation (t), temperature (t), and population (y) as independent variables. These models can help water and energy utility managers to understand energy performance better and enhance the energy efficiency of their water transfer systems.

Keywords Energy use prediction · inter-basin water transfer project · machine learning · water-energy nexus · Mokelumne River Aqueduct

✉ Sooyeon Yi
   sooyeon@berkeley.edu

1   Landscape Architecture and Environmental Planning, University of California, Berkeley, USA

2   Department of Land, Air, and Water Resources, University of California, Davis, USA

3   Energy and Resources Group, University of California, Berkeley, USA

1 Introduction

Moving and treating water are very energy-intensive processes. Approximately 30% of municipal government spending goes to energy use for drinking and wastewater (EPA, 2022). About 10% of electricity use in California accounts for moving, pumping, and treating water – much of which (4%) is used for water conveyance alone (PPIC 2018). With the rise in the water sector's electricity consumption, accurate energy consumption prediction is essential to establish plans for electricity supply and demand for water facilities (Dobschinski et al. 2017). Municipal water utilities and electricity grid operators benefit from well-predicted energy use by avoiding expensive ramping and blackouts caused by unanticipated energy use surges. Accurate energy consumption prediction can also reduce under- or over-estimating energy use (Pasha, Weathers, and Smith 2020). Overestimating energy use can lead to excessive investment in electricity supply and, ultimately, higher electricity prices (Silveira and Mata-Lima 2021). Underestimating energy use can result in electricity supply shortages, power system outages, and disruptions to water supply systems (De Felice, Alessandri, and Ruti 2013). Therefore, accurate energy use prediction is fundamental in establishing a plan for electricity supply and demand (Perelman and Fishbain 2022).

Different modeling approaches are available to predict water-related energy use. We conduct a literature review to identify the evolving approach to forecasting the water sector's energy use. Over the past decade, the prediction of energy requirements for the water-energy nexus approach in modeling has evolved to assess the water and energy systems simultaneously. As the nexus approach emerged as a frontier issue, the approach evolved into integrating the water constraints into physical-based energy models such as MARKet ALlocation/ The Integrated Markal Efom System (MARKAL/TIMES) (Suganthi and Samuel 2012). The downside of this approach is that the water system's characteristics are often not fully incorporated. This approach also makes it challenging to aggregate and synchronize the water and energy systems across different spatial and temporal scales. Another approach includes simulating the water and energy systems independently, and their outcomes pass to the other model until reaching convergence, such as Water Evaluation and Planning – Long-range Energy Alternatives Planning system (WEAP-LEAP) (Dale et al. 2015). Individual modeling evolved into integrated modeling, such as the Climate, Land, Energy, and Water framework (CLEW) (Howells et al. 2013), Water and Energy Simulation Toolset (WEST) (Thuy and Jeffers 2017), and the Water, Energy, and Food Nexus Tool (Endo et al. 2020). These models can capture the spatial variability of meteorological conditions and physical parameters within a river basin. Therefore, these models often require a large amount of topographical, soil, land use, and climate data.

In recent years, data-driven models have drawn attention to water-energy nexus modeling. The California Energy Commission, the primary energy planning agency in California, has been using regression analysis to forecast energy use in agricultural and municipal water pumping (California Energy Commission 2005). Machine learning-based models have also proven to be useful in simulating the energy use for a wastewater treatment plant (Bagherzadeh et al. 2021; Das, Kumawat, and Chaturvedi 2021; Li and Tang 2021; Zhang et al. 2021) and a distribution system (Salvino, Gomes, and Bezerra 2022). Various studies have applied machine learning algorithms in forecasting water-related energy demand but were limited to forecasting the energy use for a single water facility. Prior study has yet to examine and compare the model performance for the entire water transfer system and the subsystems

in the water transfer scheme. Machine learning algorithms may have performed well in predicting the energy use for an individual water facility. Though, little is known about predicting the energy use for a group of water facilities applying machine learning. Inter-basin water transfer projects like the State Water Project or the Mokelumne River Aqueduct may have their energy prediction models, but these models are unknown or publicly unavailable.

The novelty of this work is testing how well the machine learning-based energy forecasting model for the inter-basin water transfer project predicts the subsequent monthly energy use. As the inter-basin water transfer systems are known to be complex, with a series of dams, pumping plants, conveyance systems, treatment plants, and distribution systems, it is challenging to incorporate the physical characteristics of infrastructures and systems into a single physical-based model. The energy consumption for an inter-basin water transfer project depends on travel distance, topography, climate, water volume, pipeline length and diameter, water quality, and technology (Wakeel et al. 2016). Therefore, we attempt to demonstrate the superiority of machine learning applications in linking, analyzing, and predicting future energy use for water transfer systems. We use the multiple linear regression (MLR) as a baseline model to compare its results with three machine learning-based models (random forest (RF), deep neural network (DNN), and support vector regression (SVR)) that are non-linear. Based on the literature review, these three machine learning algorithms are well-performing models predicting energy use for a single treatment or distribution system. We use four model performance indices (coefficient of determination (R^2), root mean square error (RMSE), the mean absolute error (MAE), and percent error in peak (PEP)) to evaluate the efficiency of the baseline and machine learning-based models.

We select an inter-basin water transfer project in California. The most significant explanatory factors underlying the increasing energy consumption in California's water sector are desalination, followed by inter-basin water transfers (Sanders and Webber 2012). We choose the Mokelumne River Aqueduct system as it is less grand in scale (e.g., length, annual water delivery, cost) than the other water transfer projects in California (e.g., State Water Project, Central Valley Project).

Another novelty of this study is comparing the model performance for the entire water transfer system and its subsystems. The subsystems include conveyance (transporting water from the water source to the treatment system), treatment (moving water from the treatment plant to the distribution system), and distribution (delivering water from the distribution plant to the end-user). We exclude energy use from the other subsystems like source extraction, end-users, wastewater treatment, and recycled water, as they are outside the scope of this study. The rest of the sections are in order: Study area; Methods; Results; Discussion; Conclusions.

2 Study area

The East Bay Municipal Utility District (EBMUD) owns and operates the Mokelumne River Aqueduct, an inter-basin water transfer project providing water from the Sierra Nevada Range to urban areas in Alameda and Contra Costa counties adjacent to San Francisco Bay, California. Every year, EBMUD purchases electrical energy (for the water supply system) from six major providers including PG&E, the largest energy utility in the area.
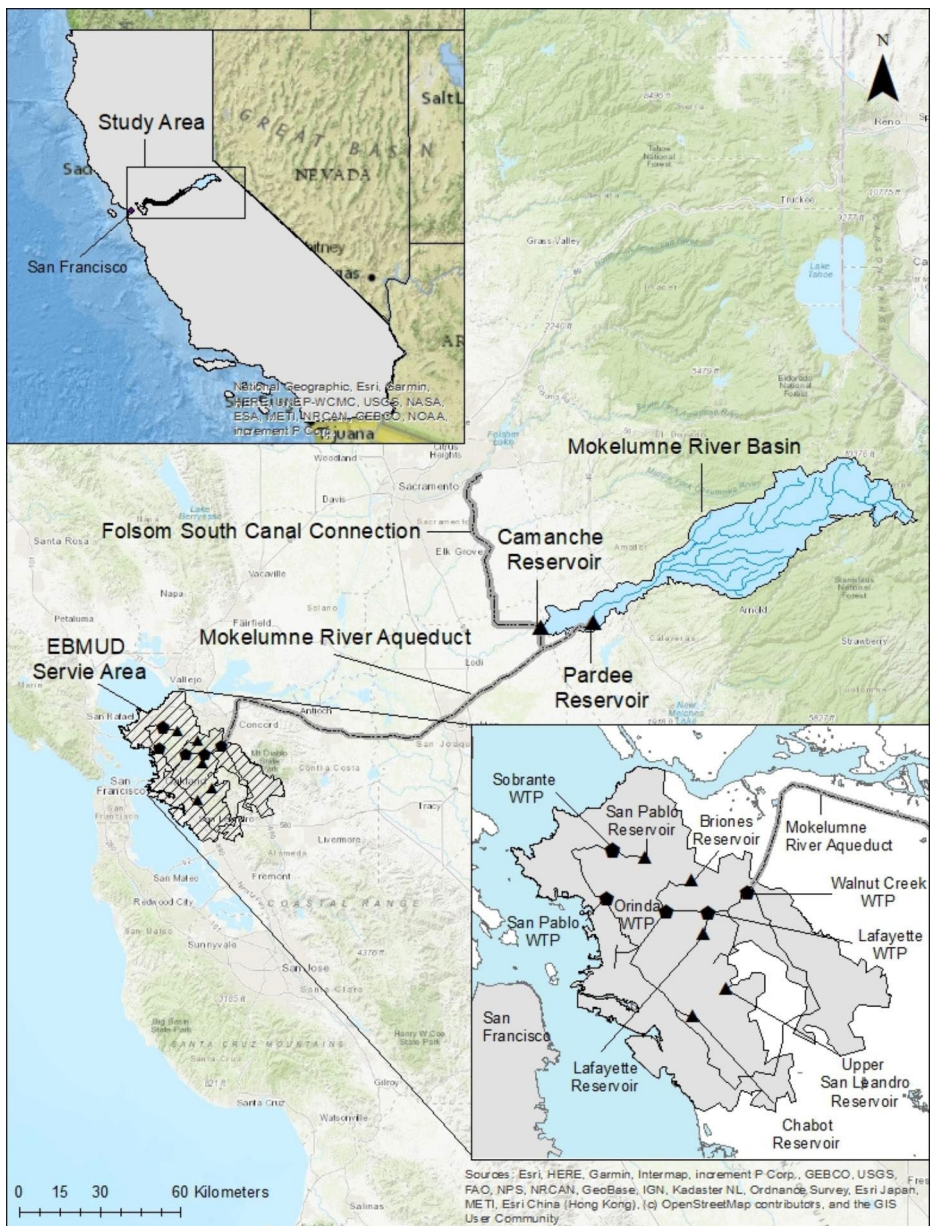
Fig. 1 Map of Mokelumne River Aqueduct system

From 1929, the Mokelumne River Aqueduct begins delivering water to East Bay cities in Alameda and Contra Costa counties (Fig. 1). Figure 1 is a map of the Mokelumne River Aqueduct System. The head waters of the Mokelumne River are in the Sierra Nevada south of Lake Tahoe. The River is largely snowmelt-fed, flowing down to the Pardee Reservoir (259million m^3 capacity) at 175m elevation. Water released from Pardee flows into Caman-

Fig. 2 The whole, conveyance, treatment, and distribution systems in the Mokelumne River Aqueduct
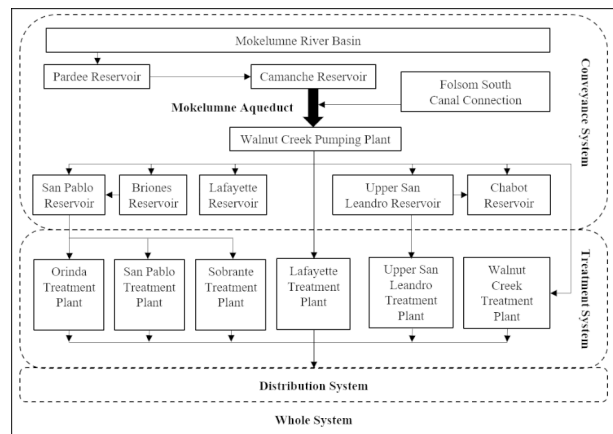
che Reservoir (532million m^3 capacity), impounded by Camanche Dam 10km downstream of Pardee Dam, operated in tandem with Pardee Reservoir. Camanche Reservoir set up to provide irrigation water, while the Pardee Reservoir provides municipal water. The Pardee and Camanche Dams have hydroelectric power facilities, though the electricity generation from these facilities is beyond the scope of this study. While the Mokelumne River Aqueduct is significantly smaller than the mega-scale inter-basin water transfer schemes in California (e.g., the California State Water Project, the Central Valley Project), it still consumes large amounts of energy for the water supply system. Where the donor basin is at high elevation, when the source is at high elevation, hydropower generation can in some cases at least partially compensate for power consumption from pumping. However, we did not consider hydropower revenues in this study, and the potential for power generation would be minor given that the elevation of Pardee Reservoir is only 175m.

The water makes a journey of over 153km across the Central Valley in pipes. The three pipelines bring the water to terminal reservoirs (San Pablo Reservoir, Briones Reservoir, Lafayette Reservoir, Upper San Leandro Reservoir, and Chabot Reservoir) or water treatment plants before it is distributed to approximately 1.4million people (Fig. 2). Figure 2 illustrates the conveyance, treatment, and distribution systems. The whole system includes energy used to pump water from the Mokelumne River Basin to end-users. We also include the energy used to move water from the Nimbus Dam on the Sacramento River through the Folsom South Canal, a water source supplementary to the main source on the Mokelumne as part of the conveyance system. The treatment system comprises energy for moving the water from some terminal reservoirs into one of the treatment plants. The distribution system is moving water from the treatment plants to the designated end-users.

EBMUD holds rights to 449million m^3 per year from the Mokelumne River. Two-thirds of the water serves residential and commercial users, and the rest for agricultural and industrial users. The Folsom South Canal Connection (completed in 2009) supplements water during droughts by drawing water from the American River near Sacramento. EBMUD uses 123,841 MWh of energy for the Mokelumne River Aqueduct at an average cost of about $13,577,872 per year. The biggest energy uses are water distribution within the service area (63%) and wastewater treatment (12%). Raw water treatment uses only 7%, conveyance from Pardee Reservoir to terminal reservoirs 4%, and unspecified other losses of 14%.

System	Dependent variable	Independent variable
Whole	Energy use for whole	Water supply for whole, precipitation, temperature, population
Conveyance	Energy use for conveyance	Water supply for conveyance, precipitation, temperature, population
Treatment	Energy use for treatment	Water supply for treatment, precipitation, temperature, population
Distribution	Energy use for distribution	Water supply for distribution, precipitation, temperature, population

Table 1 Description of dependent and independent variables for four models

3 Methods

3.1 Data Description and Feature Selection

The meteorological (e.g., precipitation, temperature, humidity, solar radiation), economical (e.g., GDP), and demographical (e.g., population) factors are the most influential elements for estimating energy consumption (Donkor et al. 2014). For all our energy use prediction models, a dependent variable is energy use, and independent variables are water supply, precipitation, temperature, and population (Table 1). Table 1 is a list of dependent and independent variables for the whole, conveyance, treatment, and distribution system. The California Data Exchange Center provides monthly precipitation data. The National Oceanic and Atmospheric Administration provides monthly temperature and relative humidity data. Annual population data comes from the US Census Bureau. The annual gross domestic product data comes from the Bureau of Economic Analysis for the Costa and Alameda Counties from 2005 to 2020. EBMUD provides monthly water supply data (1963–2020) and monthly energy use data (2005–2020). In this study, only limited water and energy data are available to work with as the overlapping water and energy data are from 2005 to 2020.

We perform the feature selection to evaluate the importance of independent variables on a dependent variable and to identify a subset of the most relevant variables for energy use prediction. We assess the importance of each variable by turning on one of the input variables while keeping the rest constant and measuring the decrease in accuracy (Breiman 2001). The initial input variables are the monthly water supply (t), water supply (t-1), water supply (t-2), water supply (t-3), water supply (t-4), temperature (t), temperature (t-1), annual population (y), precipitation (t), precipitation (t-1), annual gross domestic product (y), and relative humidity (t). Here, "t" stands for a current month, 1, 2, 3, or 4 denotes the number of months before the current month, and "y" stands for a year. The top six importance variables, or the final input variables, are water supply (t), water supply (t-1), temperature (t), population (y), precipitation (t), and precipitation (t-1).

We simulate a total of 64 cases (4 algorithms (MLR, RF, DNN, SVR) x 4 systems (whole, conveyance, treatment, distribution) x 4 scenarios (different combinations of independent variables) (Table 2). Table 2 shows four scenarios with the current monthly time step (t) and the previous monthly time step (t-1). For example, in Scenario 1 (S1), at the end of the

Table 2 Description of dependent and independent variables for three models. Monthly energy use (E), water supply (W), precipitation (R), temperature (T), and annual population (P) are the main variables in different time-steps

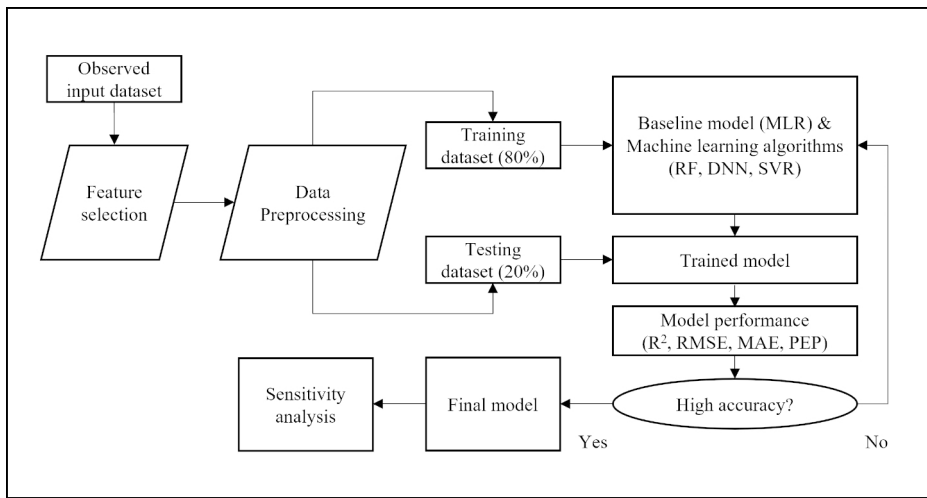| Scenario (S)	Energy use	Water supply	Precipitation	Temperature	Population
1	E (t+1)	W (t)	R (t)	T (t)	P (y)
2	E (t+1)	W(t), W(t-1)	R (t)	T (t)	P (y)
3	E (t+1)	W (t)	R (t), R (t-1)	T (t)	P (y)
4	E (t+1)	W(t), W(t-1)	R (t), R (t-1)	T (t)	P (y)

Fig. 3 Model development workflow. The workflow starts from the top-left and ends at the bottom-right

current month, we can forecast the energy use for the next month with observed data (e.g., precipitation, temperature, GDP) from the current month. Precipitation, temperature, and population are the same dataset for the four systems. Energy use and water supply datasets differ by the whole (sum of three subsystems), conveyance, treatment, and distribution systems.

3.2 Model Implementation

Figure 3 illustrates the model development procedure. We first collect the raw input dataset and perform feature selection to select the most influential independent variables on a dependent variable. We preprocess these datasets by calculating the total energy consumption for conveyance, treatment, and distribution systems. We add up the monthly energy use data from individual pumping plants. The conveyance system has nine pumping plants, the treatment system has six pumping plants, and the distribution system has 131 pumping plants. We average precipitation values from seven precipitation stations in the service area using the Thiessen polygon method in ArcGIS. These preprocessed input variables are randomly assigned to a training set (80%) and a testing set (20%). Both split sets include

randomly selected dependent and independent variables. We keep the size of training large as possible because machine learning models require a large amount of data to capture the dynamics of energy use patterns. We use 10-fold cross-validation to avoid throwing a significant portion of data away in a testing set. We evaluate the model performance based on the four model performance indices. We assess the performance of each model with R^2, RMSE, MAE, and PEP. Each performance indices for the training and testing set provides numerical values to assess the quality of training calculations and algorithm predictions during testing. We select the best machine learning model and compare it to the baseline model. We repeat this process for a total of 64 cases. The last step is performing the sensitivity analysis to determine the sensitive variables for the model. We approximate the individual variable's relative contribution to the final output's variance by squaring the rank correlation coefficient between input variables and final output and then normalizing it to 100% (Sadiq, Rajani, and Kleiner 2004). The variables with the highest relative contribution are the sensitive input variables, reducing the greatest uncertainty in the results (Hoffman and Hammonds 1992).

3.3 Machine Learning Techniques

3.3.1 Multiple Linear Regression

A baseline model is simple to set up and has a reasonable chance of providing decent results. Linear regression (LR) and multiple linear regression models often serve as baseline models for forecasting water or energy uses (Al-Musaylh et al. 2018). LR and MLR model work on any dataset size of the dataset and provides information about the relevance of the features (Kitessa et al. 2021). MLR analysis can provide decent energy demand forecasting for short- and medium-term analysis (Hong and Fan 2016) even with a limited sample size (Samuel et al. 2017). MLR is a regression model that estimates the relationship between a dependent variable and two or more independent variables. MLR is expressed as the following:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i, i = 1, 2, \ldots, n \tag{1}$$

Where Y and X denote dependent and independent variable, n is the number of datasets, β stands for the regression coefficient, p stands for the number of independent variables, and ϵ_i is the random error.

3.3.2 Random Forest

An ensemble learning technique uses the same base algorithm to provide multiple predictions and average them to produce a final model (Friedl and C.E. 1997). Random Forest (RF) is an ensemble learning technique (e.g., classification trees, regression trees) (Ho 1998). The four main features of the RF algorithm are bootstrap resampling, random feature selection, out-of-bag error estimation, and a fully grown decision tree (Jiang et al. 2009). The parameterization of RF is much simpler and computationally lighter than other machine learning algorithms. The number of trees in the forest (n_{tree}) and the number of random variables in each tree (m_{try}) are the two main parameters (Breiman et al. 2018). RF also offers

a function of assessing the relative importance of the features (variables), and selecting the most important features, and reducing dimensionality (Breiman 2001).

3.3.3 Deep Neural Network

Neural networks have extensive applications in predicting behaviors (Goodfellow, Bengio, and Courville 2016). The neural network is popular for short- and medium-term energy forecasting (Azadeh et al. 2007). An artificial neural network (ANN) uses one hidden layer (e.g., one layer between input and output). A deep neural network is an ANN with multiple hidden layers between the input and output layers. These extra layers enable the composition of features from lower layers, potentially modeling complex data with fewer units than similarly performing networks (Goodfellow, Bengio, and Courville 2016). The main parameters are the number of layers, number of neurons per layer, and number of training iterations. Normalizing the input dataset for DNN is a well-established technique for improving the convergence properties of a network. We normalize the input datasets using a max-min normalization technique. The mathematical formation is as follows:

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{2}$$

3.3.4 Support Vector Regression

The support vector algorithm (SVA) is for a nonlinear generalization algorithm (Vapnik 1963). The support vector regression originates from SVA (Cortes 1995). SVR is the nonlinear mapping of the original data into a high-dimensional feature space using a kernel function (Kalra, Ahmad, and Nayak 2013). The kernel function includes linear, polynomial, sigmoid, and radial bases (Gaussian) (Boser, Guyon, and Vapnik 1992). We choose the radial basis function, the most common kernel function applied in hydrology (Dibike et al. 2001). SVR model shows good performance in medium-term energy forecasting (He et al. 2017). SVM is an optimal method for regression problems with a small dataset (Minghui and Chuanfeng 2015). We normalize the input datasets using a max-min normalization technique. The regularization parameter (C), the tolerance threshold (ε), and the width of the radial basis function (ϒ) are the key parameters.

3.4 Performance Indices

We compute four model performance indices, which are a coefficient of determination (R^2), root mean square error (RMSE), the mean absolute error (MAE), and percent error in peak (PEP). These indices are well-known for evaluating the reliability of hydrological models (Doycheva et al. 2017). R^2 represents the fitness of observed and predicted data on 45 degrees reference line (Kazemi and Barati 2022). A higher coefficient indicates a better fit for the model. The R^2 is the following:

$$R^2 = \frac{1 - \sum_{i=1}^{n} (y_i - f_i)^2}{\sum_{i=1}^{n} (y_i - \widehat{y})^2} \tag{3}$$

Where n is the total number of observed data, y_i is the observed value for data point i, $\widehat{y}$ is the mean of the observed data, and f_i is the predicted value for data point i. RMSE is the square root of the mean square of all errors (Alizadeh et al. 2017). It measures the error of a model in predicting quantitative data. A small RMSE value indicates that the model predicts the observed data well, while a large RMSE value generally means that the model fails to account for key features underlying the data (Sadeghifar and Barati 2018). RMSE is the following:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2} \tag{4}$$

MAE represents the average absolute difference between the observed and predicted values in the data (Shu et al. 2022). It measures the average magnitude of the error. A lower MAE indicates a better model, and 0 means the model is perfect. MAE is the following:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - f_i| \tag{5}$$

The percent error in peak energy use (PEP) is the observed and simulated peak flow rate ratio. It can be expressed as a percentage error in the simulated peak (Green and Stephenson 1986). The high PEP indicates that a model is good at forecasting peak values. PEP is the following:

$$PEP = \frac{q_{ps} - q_{po}}{q_{po}} * 100 \tag{6}$$

q_{ps} is simulated peak flow rate and q_{po} is an observed peak flow rate.

4 Results

4.1 Whole System

The whole system evaluated the energy use of transporting water from the conveyance system to the treatment system to the distribution system. RF model had the highest R^2 (0.677), while the DNN model had the lowest RMSE (1,386,206), MAE (952,885), and PEP (-8.30%) in the testing set. DNN was the best performing model as DNN had the three best model performance indices (RMSE, MAE, PEP), while RF had the best performance index (R^2). The combination of DNN and Scenario 4 was the best performing case for the whole system (Table 3). Table 3 displays the performance indices for the whole system from Scenario 1 to Scenario 4 for training and testing sets.

Table 3 The performance indices for the whole system for training and testing sets

Machine Learning	Scenario	Training				Testing			
		R^2	RMSE	MAE	PEP(%)	R^2	RMSE	MAE	PEP(%)
MLR	**S1**	**0.450**	**1,542,548**	**1,106,218**	**-30.6**	**0.557**	**1,336,463**	**922,091**	**-33.8**
	S2	0.495	1,378,833	987,750	-31.6	0.508	1,768,801	1,216,824	-35.5
	S3	0.470	1,413,648	1,010,382	-31.9	0.486	1,806,431	1,255,409	-37.8
	S4	0.497	1,376,990	987,670	-31.5	0.506	1,770,393	1,204,245	-35.2
RF	S1	0.771	1,075,445	839,576	-20.0	0.575	1,336,597	1,031,332	-18.1
	S2	**0.841**	**831,638**	**667,930**	**-14.5**	**0.677**	**1,416,202**	**982,208**	**-11.8**
	S3	0.839	845,990	664,513	-17.1	0.639	1,515,880	1,093,049	-19.0
	S4	0.845	829,874	641,899	-19.3	0.664	1,464,269	1,012,304	-20.7
DNN	S1	0.817	1,092,233	670,472	-4.4	0.500	1,567,506	952,088	-11.9
	S2	0.829	1,055,545	697,790	-28.4	0.515	1,543,475	976,098	-14.3
	S3	0.840	1,022,153	698,634	-30.0	0.535	1,510,419	928,367	-13.1
	S4	**0.938**	**638,435**	**392,481**	**-21.9**	**0.609**	**1,386,206**	**952,885**	**-8.3**
SVR	**S1**	**0.509**	**1,529,373**	**943,969**	**-29.5**	**0.630**	**1,162,163**	**924,879**	**-28.2**
	S2	0.509	1,469,970	878,375	-32.9	0.415	1,697,739	1,041,907	-35.7
	S3	0.473	1,643,877	1,353,171	-26.8	0.514	1,343,831	1,140,952	-0.9
	S4	0.484	1,657,578	1,392,292	-30.7	0.526	1,374,669	1,206,161	-26.9

4.2 Conveyance System

The main challenge for the conveyance system was capturing and predicting irregular energy use patterns in two cases. The first occasion was when the Camanche Reservoir released water to fill the low water levels in the terminal reservoirs. Another occasion was when the Folsom South Canal Connection released water to service areas during drought. Despite these two irregularities that made the energy use pattern inconsistent, the DNN model showed outstanding model performance. Among the three machine learning algorithms, the RF model showed the highest R^2 (0.774), while the DNN model had the lowest RMSE (757,556), MAE (359,303), and PEP (15.9%) in the testing set. The combination of DNN with Scenario 2 was the best-performing case for the conveyance system (Table 4). Table 4 shows the performance indices for the conveyance system from Scenario 1 to Scenario 4 for training and testing sets.

4.3 Treatment System

Among the three machine learning algorithms, the SVR model showed the highest R^2 (0.859) and PEP (-3.8%), while the DNN model had the lowest RMSE (169,694) and MAE (126,722) in the testing set. The SVR model had the two best model performance indices (R^2, PEP), and the DNN model also had the two best model performance indices (RMSE, MAE). The difference in R^2 between DNN and SVR was 1.16%, which is marginal. Therefore, the combination of DNN with Scenario 4 was the best performing case for the treatment system (Table 5). Table 5 shows the performance indices for the treatment system from Scenario 1 to Scenario 4 for training and testing sets.

Table 4 The performance indices for the conveyance system for training and testing sets

Conveyance system

Machine Learning	Scenario	Training				Testing			
		R^2	RMSE	MAE	PEP(%)	R^2	RMSE	MAE	PEP(%)
MLR	S1	0.631	1,006,638	701,351	-28.7	0.617	1,335,128	815,752	-30.4
	S2	0.566	1,065,974	727,338	-33.2	0.751	1,141,638	793,532	-33.7
	S3	**0.562**	**1,071,553**	**728,372**	**-28.3**	**0.764**	**1,090,427**	**774,177**	**-30.4**
	S4	0.573	1,057,454	723,333	-31.2	0.743	1,142,990	808,140	-32.9
RF	S1	0.776	845,348	643,872	-27.1	0.463	1,590,625	1,073,138	-39.5
	S2	**0.767**	**815,413**	**598,812**	**-25.8**	**0.774**	**1,045,818**	**825,818**	**-21.4**
	S3	0.734	924,438	683,514	-39.1	0.757	1,181,645	958,082	-36.6
	S4	0.712	877,065	623,238	-32.5	0.749	1,095,412	792,981	-27.8
DNN	S1	0.862	576,774	280,335	-5.6	0.707	776,012	375,018	-10.6
	S2	**0.874**	**550,887**	**255,573**	**-8.7**	**0.721**	**757,556**	**359,303**	**15.9**
	S3	0.870	559,799	276,037	-9.3	0.698	788,057	372,133	18.6
	S4	0.865	570,900	281,002	-0.8	0.719	760,264	342,999	-28.9
SVR	**S1**	**0.674**	**1,005,512**	**605,470**	**-30.6**	**0.716**	**1,036,197**	**662,740**	**-23.0**
	S2	0.774	808,289	456,909	-26.3	0.294	1,771,161	990,931	-20.0
	S3	0.743	861,344	519,592	-23.1	0.541	1,528,800	873,585	-23.2
	S4	0.764	833,225	482,861	-16.0	0.332	1,780,727	952,618	-65.2

Table 5 The performance indices for the treatment system for training and testing sets

Treatment system

Machine Learning	Scenario	Training				Testing			
		R^2	RMSE	MAE	PEP(%)	R^2	RMSE	MAE	PEP(%)
MLR	S1	0.549	296,691	235,770	-16.2	0.740	305,639	247,181	-12.2
	S2	**0.647**	**262,547**	**213,108**	**-13.8**	**0.738**	**305,291**	**235,694**	**-11.1**
	S3	0.596	280,679	224,139	-15.2	0.709	318,555	254,135	-10.5
	S4	0.648	261,922	212,945	-13.8	0.732	307,976	237,076	-11.1
RF	S1	0.836	186,857	148,615	-12.7	0.797	277,061	221,547	-21.2
	S2	**0.874**	**168,097**	**133,007**	**-11.6**	**0.818**	**280,008**	**220,774**	**-22.4**
	S3	0.866	171,980	135,257	-11.7	0.803	283,706	226,725	-21.6
	S4	0.896	150,409	119,639	-11.6	0.768	286,436	228,389	-23.4
DNN	S1	0.836	270,152	182,933	-8.2	0.653	257,073	178,929	14.2
	S2	0.942	160,742	106,418	-4.8	0.779	205,363	142,608	-3.6
	S3	0.928	178,583	126,691	-5.4	0.753	217,061	138,452	-5.0
	S4	**0.940**	**162,843**	**108,884**	**0.7**	**0.849**	**169,694**	**126,722**	**-7.3**
SVR	S1	0.805	204,589	158,682	-1.6	0.745	261,142	196,456	-5.1
	S2	**0.885**	**155,150**	**112,473**	**-3.1**	**0.859**	**207,019**	**177,179**	**-3.8**
	S3	0.848	171,832	129,387	-1.5	0.789	269,818	186,942	-4.8
	S4	0.860	165,334	127,682	-3.2	0.782	270,500	195,075	-1.3

4.4 Distribution System

The distribution system had a total of 130 pumping plants, a 20-fold greater density of pumping plants than the conveyance and treatment systems. This system was more vulnerable to uncertainties and errors. Among the three machine learning algorithms, the DNN

Table 6 The performance indices for the distribution system for training and testing sets

Distribution system									
Machine Learning	Scenario	Training				Testing			
		R^2	RMSE	MAE	PEP(%)	R^2	RMSE	MAE	PEP(%)
MLR	**S1**	**0.758**	**508,255**	**412,320**	**-4.08**	**0.829**	**478,915**	**387,446**	**-10.80**
	S2	0.848	407,812	326,029	-5.81	0.771	531,223	404,037	3.12
	S3	0.797	471,650	383,173	-6.23	0.734	574,512	440,748	0.99
	S4	0.850	404,782	327,375	-5.81	0.762	542,155	424,537	3.18
RF	S1	0.832	431,241	348,244	-13.02	0.792	544,544	440,915	-18.35
	S2	0.896	350,943	288,305	-8.41	0.762	552,179	403,528	-4.38
	S3	0.892	357,541	297,113	-9.56	0.743	570,999	426,118	-4.30
	S4	**0.920**	**302,493**	**246,795**	**-8.59**	**0.784**	**519,199**	**399,027**	**-3.33**
DNN	S1	0.888	427,071	311,739	-1.02	0.878	504,953	357,772	-14.96
	S2	0.927	345,748	241,590	-4.84	0.908	438,147	296,591	-7.14
	S3	0.899	406,034	282,695	-5.44	0.936	366,610	271,910	-7.47
	S4	**0.943**	**305,107**	**201,265**	**0.72**	**0.923**	**401,397**	**269,891**	**-4.86**
SVR	S1	0.781	479,212	366,382	-8.46	0.866	484,672	409,624	-8.29
	S2	**0.881**	**358,390**	**266,225**	**-6.72**	**0.894**	**386,120**	**318,847**	**-9.88**
	S3	0.837	445,347	319,366	-9.59	0.829	398,512	330,269	-8.91
	S4	0.871	396,678	293,254	-6.84	0.872	345,424	282,936	-9.23

model showed the highest R^2 (0.923) and MAE (269,891), while the SVR model had the lowest RMSE (386,120), and the RF model had the lowest PEP (-3.33%) in the testing set. PEP was a less important factor as the energy use forecasting models' primary goal was to predict the overall energy use, not the peak energy use. The combination of DNN with Scenario 4 was the best-performing case for the distribution system (Table 6). Table 6 shows the performance indices for the distribution system from Scenario 1 to Scenario 4 for training and testing sets.

4.5 Sensitivity Analysis

Sensitivity analysis investigated the sensitivity of input variables (water supply (t), water supply (t-1), temperature (t), population (y), precipitation (t), and precipitation (t-1)) to the variations in the final outputs. The water supply (t) (24.9%) and water supply (t-1) (9.28%) were the most sensitive independent variables in predicting energy use (dependent variable). Temperature (t) (5.24%), population (t) (4.71%), and precipitation (t-1) were the next sensitive variables in order.

5 Discussion

The essential step in the water-energy nexus approach in modeling was to compare the results from this study to similar research from the past literature. Having limited similar research made it challenging to compare the results from this study to those of other studies. We found no comparative research on predicting energy use for the whole or the conveyance system. Prior studies have applied machine learning algorithms to simulate the energy use for wastewater treatment plants (Bagherzadeh et al. 2021; Das, Kumawat, and Chaturvedi

2021; Li and Tang 2021; Zhang et al. 2021) and distribution systems (Antonopoulos and Gianniou 2022; Salvino, Gomes, and Bezerra 2022). Prior studies concluded that the ANN model was useful in simulating the energy efficiency of the water distribution system (Das, Kumawat, and Chaturvedi 2021; Salvino, Gomes, and Bezerra 2022). We also found that the DNN model, an ANN model with multiple hidden layers between the input and output layers, was the most effective machine learning algorithm in predicting the energy use of the water distribution system. Other studies found that the RF model effectively predicted energy consumption for wastewater treatment plants (Zhang et al. 2021). We concluded that DNN and SVR models performed better than the RF model for the treatment system (in the inter-basin water transfer project), which was not the same as the wastewater treatment plant.

The machine learning-based models (RF, DNN, and SVR) generated better model performance than the baseline model (MLR) except for RF models in the distribution system. In the distribution system, MLR model performed 5.43% (R^2), 7.76% (RMSE), 2.90% (MAE) better and 69.1% (PEP) worse than RF model. The difference was marginal, and we still concluded that machine learning-based models performed better than the baseline model (Figs. 4 and 5). Figure 4 is a scatter plot of the observed and predicted energy use on a 45-degree reference line. We plotted training and testing sets for the whole and the subsystems. However, these scatter plots did not show the model's performance in predicting the overall energy use pattern or the peak energy use. To address this limitation, we plotted observed and predicted energy use for the testing set over 30 months (Fig. 5). MLR was a simpler model that predicted the dependent variable well and did not require much expertise and time to build and analyze. Developing such energy prediction models through multiple linear modeling techniques caused certain disadvantages since the behavior of certain independent variables for energy consumption was non-linear. Thus, forecasting the energy use for the water transfer system through non-linear modeling (RF, DNN, SVR) yielded higher model performance indices than linear modeling (MLR).

Among the three non-linear modelings, DNN performed better than RF and SVR. Among the four scenarios, Scenario 4 yielded the best in DNN models. A combination of DNN and Scenario 4 was the best case for the whole, treatment, and distribution systems except for the conveyance system. In the conveyance system, Scenario 2 performed 0.028% (R^2), 0.35% (RMSE), 4.54% (MAE) and 45% (PEP) better than Scenario 4. Given that the differences in R2, RMSE, and MAE between Scenario 2 and Scenario 4 were marginal, we concluded that DNN and Scenario 4 were the best combinations for predicting the energy use for the Mokelumne River Aqueduct.

The subsystems (conveyance, treatment, and distribution) showed better model performance than the whole system. Coincidentally, the energy use increase in one subsystem and decrease in another could offset the characteristics and not reflect the model's results. Subsystems reflected their unique energy use characteristics and produced better model performance than a whole system. Therefore, predicting the energy use for the subsystems was more economically efficient than for the whole system. EBMUD decision-makers and operators can predict the energy use for each subsystem and offer energy savings in either conveyance, treatment, or distribution systems.

There were some limitations and uncertainties in this study. The overlapping period of the water supply and energy use dataset was relatively short. Models training on a small number of observations were vulnerable to overfitting and producing inaccurate results. The
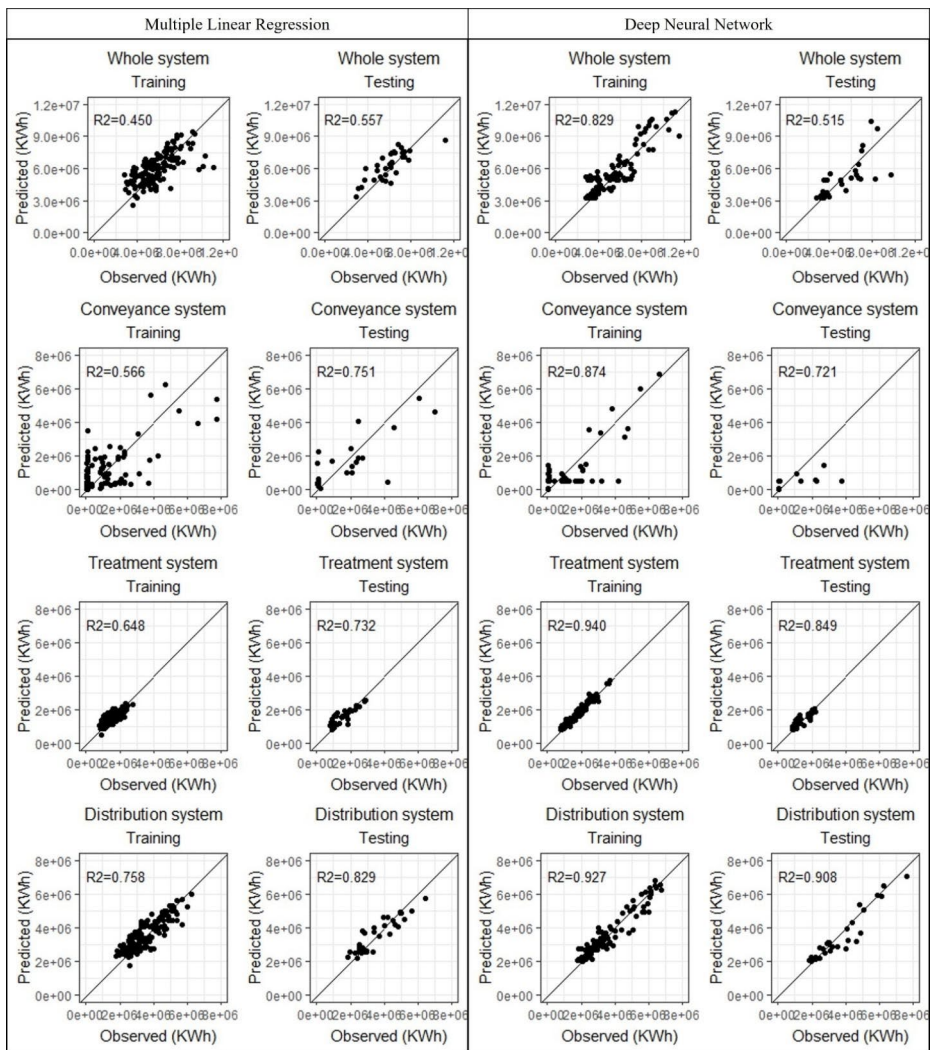
Fig. 4 The observed and predicted energy use are plotted against a 45-degree reference line. The first row is the whole system, the second row is the conveyance system, the third row is the treatment system, and the fourth row is the distribution system. The first two columns are the results of training and testing sets from the baseline model (MLR). The third and fourth columns are the results of training and testing sets from the best-performing model (DNN).

results showed that R^2 is 19.2%, 26.4%, 12.3%, and 5.26% higher in the testing set than the training set in MLR for the whole, conveyance, treatment, and distribution system, respectively. A small dataset or the split of training and testing sets were possible reasons behind this. The results presented that R^2 is 19.2%, 5.87%, and 1.45% higher in the testing set than the training set in SVR for the whole, conveyance, and distribution system, respectively. We used a relatively simple baseline model and machine learning models with fewer parameters to tune to address this limitation and avoid overfitting. We used 10-fold cross-validation to
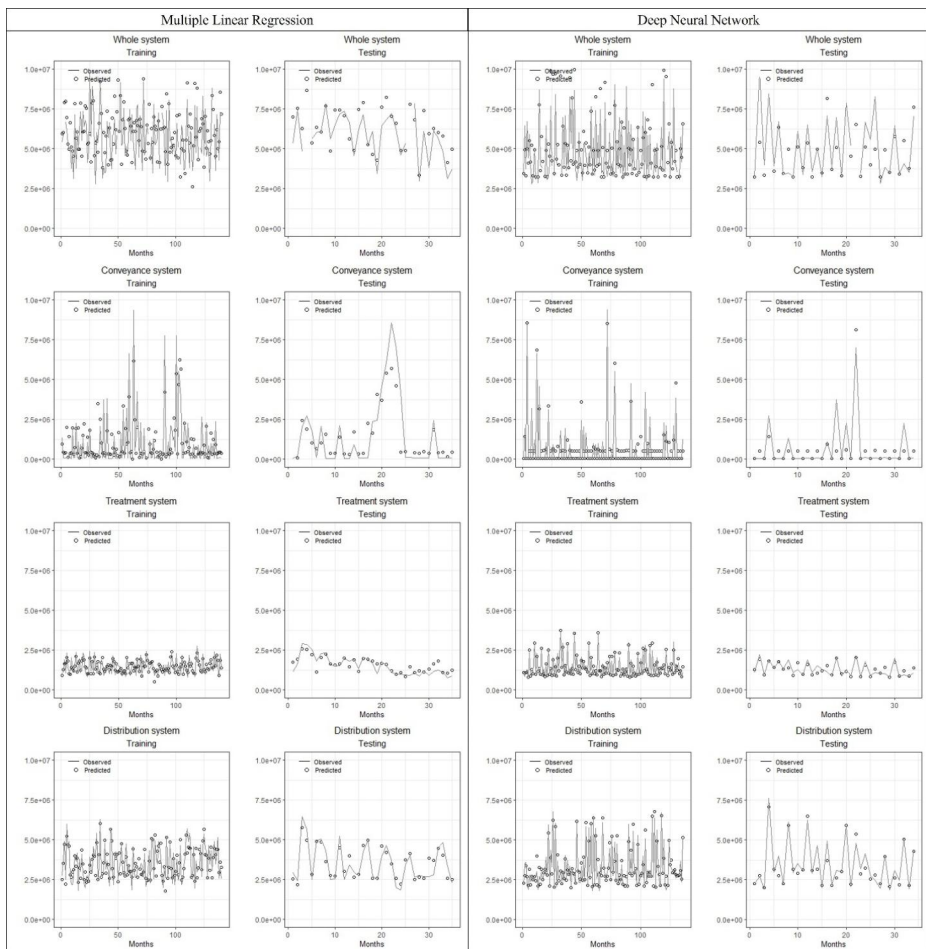
Fig. 5 The observed and predicted energy use are plotted for testing set over 30 months. These months do not indicate a specific time of the month. The first row is the whole system, the second row is conveyance system, the third row is the treatment system, and the fourth row is the distribution system. The first two columns are the results of training and testing sets from the baseline model (MLR). The third and fourth columns are the results of training and testing sets from the best-performing model (DNN).

avoid underfitting as models use all the data for training or testing and overfitting as all the datasets are used in the validation set.

6 Conclusion

This study demonstrated the effectiveness of applying machine learning algorithms to build energy use forecasting models for the inter-basin water transfer system and its subsystems. RF, DNN, and SVR (non-linear) models were better at reflecting the non-linear characteristics and energy use patterns in the water transfer system than the MLR model (linear). DNN models yielded higher model performance indices than RF and SVR models. Subsystems

well reflected their unique energy use characteristics and avoided offset from subsystems which eventually yielded better model performance than a whole system. The water supply (t) and water supply (t-1) were the most sensitive parameters. The best case was having water supply (t), water supply (t-1), precipitation (t), temperature (t), and population (y) as independent variables to predict energy use, which is a dependent variable. These developed models can support water and energy utility managers and planners to understand energy performance and enhance energy efficiency for inter-basin water transfer projects. Specifically, EBMUD decision-makers and operators can use these models to predict the monthly energy consumption with projected water supply, precipitation, temperature, and population dataset.

Potential future studies include testing different lead times and time resolutions. For example, projecting hourly or daily energy use can capture abrupt changes in variables. Future studies include developing energy consumption forecasting models using different machine learning algorithms for the Mokelumne River Aqueduct or the other inter-basin water transfer projects. The Mokelumne River Aqueduct is a case study in this paper to illustrate the model by combining a specific jurisdiction with diversity in energy resources and deep concerns about water availability.

Forecasting the future energy use behaviors and patterns are essential in proposing energy-efficient planning that saves cost, reduces operational risks, and supports better decision-making. Energy use forecasting for inter-basin water transfer projects is crucial in balancing the electricity supply and demand. Energy use predictions support reliable water and energy grid operation and encourage cost-effective operation by improving generation scheduling. This study serves as an excellent model for future studies that attempt to forecast the energy consumption for other energy-intensive water systems and can expand for more complex water transfer projects. As EBMUD will benefit from these models in understanding energy use for one of the most populous counties in the Bay Area, other municipalities and governments can replicate these methods to address energy use in a time of rapid global warming, population growth, and economic change.

Author Contribution Sooyeon Yi: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing Original Draft, Visualization. G. Mathias Kondolf: Supervision, Validation, Writing Review & Editing. Samuel Sandoval Solis: Supervision, Validation, Writing Review & Editing. Larry Dale: Supervision, Validation, Writing Review & Editing.

Data Availability Some data are available from the corresponding author upon requests.

Declarations

Statements and Declarations Ethical Approval: Not applicable.

Consent to Participate: Not applicable.

References

Al-Musaylh MS, Ravinesh C, Deo, Jan F, Adamowski, Li Y (2018) Short-Term Electricity Demand Forecasting with MARS, SVR and ARIMA Models Using Aggregated Demand Data in Queensland, Australia. Adv Eng Inform 35:1–16. https://doi.org/10.1016/j.aei.2017.11.002

Alizadeh MJ et al (2017) Prediction of Longitudinal Dispersion Coefficient in Natural Rivers Using a Cluster-Based Bayesian Network. Environ Earth Sci 76(2):86. https://doi.org/10.1007/s12665-016-6379-6

Antonopoulos VZ, Gianniou SK (2022) Analysis and Modelling of Temperature at the Water – Atmosphere Interface of a Lake by Energy Budget and ANNs Models. Environ Processes 9(1):15. https://doi.org/10.1007/s40710-022-00572-0

Azadeh A, Ghaderi SF, Sohrabkhani S (2007) Forecasting Electrical Consumption by Integration of Neural Network, Time Series and ANOVA. Appl Math Comput 186(2):1753–1761. https://doi.org/10.1016/j.amc.2006.08.094

Bagherzadeh F, Nouri AS, Mehrani MJ, and Suresh Thennadil (2021) Prediction of Energy Consumption and Evaluation of Affecting Factors in a Full-Scale WWTP Using a Machine Learning Approach. Process Saf Environ Prot 154:458–466. https://doi.org/10.1016/j.psep.2021.08.040

Boser BE, Isabelle M, Guyon, Vapnik VN (1992) "A Training Algorithm for Optimal Margin Classifiers." *Proceedings of the fifth annual workshop on Computational learning theory – COLT '92.*: 144-152. https://doi.org/10.1145/130385.130401

Breiman L (2001) Random Forest. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Breiman L, Cutler A, Liaw A, Liaw A (2018) *Breiman and Cutler's Random Forests for Classification and Regression*

California Energy Commission (2005) *Energy Demand Forecast Methods Report*.

Cortes C, Vapnik V (1995) Support-Vector Networks. Mach Learn 20(273):297. https://doi.org/10.1007/BF00994018

Dale LL et al (2015) An Integrated Assessment of Water-Energy and Climate Change in Sacramento, California: How Strong Is the Nexus? Clim Change 132(2):223–235. https://doi.org/10.1007/s10584-015-1370-x

Das A, Kumawat PK, and Chaturvedi ND (2021) A study to target energy consumption in wastewater treatment plant using machine learning algorithms (pp. 1511–1516). https://doi.org/10.1016/B978-0-323-88506-5.50233-3

Dibike YB, Velickov S, Solomatine D, Abbott MB (2001) Model Induction with Support Vector Machines: Introduction and Applications. J Comput Civil Eng 15(3):208–216. https://doi.org/10.1061/(ASCE)0887-3801(2001)15:3(208)

Dobschinski J et al (2017) Uncertainty Forecasting in a Nutshell: Prediction Models Designed to Prevent Significant Errors. IEEE Power Energ Mag 15(6):40–49. https://doi.org/10.1109/MPE.2017.2729100

Donkor EA et al (2014) Urban Water Demand Forecasting: Review of Methods and Models. Journal of Water Resources Planning and Management, 140(2), 146–159. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000314

Doycheva K et al(2017) Assessment and weighting of meteorological ensemble forecast members based on supervised machine learning with application to runoff simulations and flood warning. Advanced Engineering Informatics, 33, 427–439. https://doi.org/10.1016/j.aei.2016.11.001

Endo A et al (2020) Dynamics of Water–Energy–Food Nexus Methodology, Methods, and Tools. Curr Opin Environ Sci Health 13:46–60. https://doi.org/10.1016/j.coesh.2019.10.004

De Felice, Matteo A, Alessandri, and Paolo M. Ruti (2013) Electricity Demand Forecasting over Italy: Potential Benefits Using Numerical Weather Prediction Models. Electr Power Syst Res 104:71–79. https://doi.org/10.1016/j.epsr.2013.06.004

Friedl MA, Broadley CE(1997) Decision Tree Classification of Land Cover from Remotely Sensed Data. Remote Sensing of Environment, 61(3), 399–409. https://doi.org/10.1016/S0034-4257(97)00049-7

Goodfellow I, Bengio Y, and Aaron Courville (2016) Deep Learning. The MIT Press

Green IRA, Stephenson D (1986) Criteria for Comparison of Single Event Models. Hydrol Sci J 31(3):395–411. https://doi.org/10.1080/02626668609491056

He Y et al (2017) Urban Long Term Electricity Demand Forecast Method Based on System Dynamics of the New Economic Normal: The Case of Tianjin. Energy 133:9–22. https://doi.org/10.1016/j.energy.2017.05.107

Ho TK(1998) The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), 832–844. https://doi.org/10.1109/34.709601

Hammonds, JS, Hoffman, FO, & Bartell, SM(1994) An introductory guide to uncertainty analysis in environmental and health risk assessment. Environmental Restoration Program. https://doi.org/10.2172/10127301

Hong T, and Shu Fan (2016) Probabilistic Electric Load Forecasting: A Tutorial Review. Int J Forecast 32(3):914–938. https://doi.org/10.1016/j.ijforecast.2015.11.011

Howells M et al (2013) Integrated Analysis of Climate Change, Land-Use, Energy and Water Strategies. Nat Clim Change 3(7):621–626. https://doi.org/10.1038/nclimate1789

Jiang R, Tang W, Wu X, Fu W(2009) A random forest approach to the detection of epistatic interactions in case-control studies. BMC Bioinformatics, 10. https://doi.org/10.1186/1471-2105-10-S1-S65

Kalra A, Sajjad Ahmad, and Anurag Nayak (2013) Increasing Streamflow Forecast Lead Time for Snowmelt-Driven Catchment Based on Large-Scale Climate Patterns. Adv Water Resour 53:150–162. https://doi.org/10.1016/j.advwatres.2012.11.003

Kazemi M, and Reza Barati (2022) Application of Dimensional Analysis and Multi-Gene Genetic Programming to Predict the Performance of Tunnel Boring Machines. Appl Soft Comput 124:1568-4946. https://doi.org/10.1016/j.asoc.2022.108997

Kitessa B, Dessalegn SM, Ayalew GS, Gebrie, and Solomon Tesfamariam Teferi (2021) Long-Term Water-Energy Demand Prediction Using a Regression Model: A Case Study of Addis Ababa City. J Water Clim Change 12(6):2555–2578. https://doi.org/10.2166/wcc.2021.012

Li J, Tang WZ (2021) Improved Unit Energy Efficiency and Reduced Cost by Innovative Industrial Wastewater Treatment Systems. Environ Processes 8(4):1433–1454. https://doi.org/10.1007/s40710-021-00544-w

Minghui M, and Zhao Chuanfeng (2015) Application of Support Vector Machines to a Small-Sample Prediction. Adv Petroleum Explor Dev 10(2):72–75. https://doi.org/10.3968/7830

Pasha M, Fayzul K, Weathers M, and Brennan Smith (2020) Investigating Energy Flow in Water-Energy Storage for Hydropower Generation in Water Distribution Systems. Water Resour Manage 34(5):1623–1623. https://doi.org/10.1007/s11269-020-02539-y

Perelman G, and Barak Fishbain (2022) Critical Elements Analysis of Water Supply Systems to Improve Energy Efficiency in Failure Scenarios. Water Resour Manage 36(10), 3797–3811. https://doi.org/10.1007/s11269-022-03232-y

PPIC (2018) *Energy and Water*. San Francisco. https://www.ppic.org/wp-content/uploads/californias-water-energy-and-water-november-2018.pdf

Sadeghifar T, Barati R (2018) Application of Adaptive Neuro-Fuzzy Inference System to Estimate Alongshore Sediment Transport Rate (A Real Case Study: Southern Shorelines of Caspian Sea). J Soft Comput Civil Eng 2(4):72–85. https://doi.org/10.22115/SCCE.2018.135975.1074

Sadiq R, Rajani B, and Yehuda Kleiner (2004) Probabilistic Risk Analysis of Corrosion Associated Failures in Cast Iron Water Mains. Reliab Eng Syst Saf 86(1):1–10. https://doi.org/10.1016/j.ress.2003.12.007

Salvino L, Régis HP, Gomes, Saulo de Tarso Marques Bezerra (2022) Design of a Control System Using an Artificial Neural Network to Optimize the Energy Efficiency of Water Distribution Systems. Water Resour Manage 36(8):2779–2793. https://doi.org/10.1007/s11269-022-03175-4

Samuel IA et al (2017) A Vomparative Dtudy of Tegression Analysis and Artificial Neural Network Methods for Medium-Term Load Forecasting. Article in Indian Journal of Science and Technology 10(10):974–6846. https://doi.org/10.17485/ijst/2017/v10i10/86243

Sanders KT, Webber ME(2012) Evaluating the energy consumed for water use in the United States. Environmental Research Letters, 7(3). https://doi.org/10.1088/1748-9326/7/3/034034

Shu X et al(2022) Multi-Step-Ahead Monthly Streamflow Forecasting Using Convolutional Neural Networks. Water Resources Management, 36(11), 3949–3964. https://doi.org/10.1007/s11269-022-03165-6

Silveira APP(2021) Assessing Energy Efficiency in Water Utilities Using Long-term Data Analysis. Water Resources Management, 35(9), 2763–2779. https://doi.org/10.1007/s11269-021-02866-8

Suganthi L, Samuel AA (2012) Energy Models for Demand Forecasting - A Review. Renew Sustain Energy Rev 16(2):1223–1240. https://doi.org/10.1016/j.rser.2011.08.014

Thuy N, and Robert Jeffers (2017) "Water Energy Simulation Toolset. " Idaho National Laboratory

Vapnik V, Chervonenkis (1963) Pattern Recognition Using Generalized Portrait Method. Autom Remote Control 24:774–780

Wakeel M et al (2016) Energy Consumption for Water Use Cycles in Different Countries: A Review. Appl Energy 178(19):868–885. https://doi.org/10.1016/j.apenergy.2016.06.114

Zhang S, Wang H, Keller AA (2021) Novel Machine Learning-Based Energy Consumption Model of Wastewater Treatment Plants. ACS ES&T Water 1(12):2531–2540. https://doi.org/10.1021/acsestwater.1c00283