

Prediction of multi-sectoral longitudinal water withdrawals using hierarchical machine learning models

Julie Shortridge Department of Biological Systems Engineering, Virginia Tech, 155 Ag Quad Lane, Blacksburg, Virginia 24061, USA
E-mail: jshortridge@vt.edu JS, 0000-0003-1612-5740

ABSTRACT

Accurate models of water withdrawal are crucial in anticipating the potential water use impacts of drought and climate change. Machine learning methods can simulate the complex, nonlinear relationship between water use and potential explanatory factors, but rarely incorporate the hierarchical nature of water use data. This work presents a novel approach for the prediction of water withdrawals across multiple usage sectors using an ensemble of models fit at different hierarchical levels. Models were fit at the facility and sectoral grouping levels, as well as across facility clusters defined by temporal water use characteristics. Using repeated holdout cross-validation and a dataset of over 300,000 observations of monthly water withdrawal across 1,509 facilities, it demonstrates that ensemble predictions led to statistically significant improvements in predictive performance in five of the eight sectors analyzed. The use of ensemble modeling resulted in lower predictive errors compared to facility models in 65% of facilities analyzed. The relative improvement gained by ensemble modeling was greatest for facilities with fewer observations and higher variance, indicating its potential value in predicting withdrawal for facilities with relatively short data records or data quality issues.

Key words: hierarchical modeling, machine learning, water demand, water resources management

HIGHLIGHTS

- Hierarchical ensemble models reduce predictive errors for a majority of facilities analyzed.
- Cluster analysis is used to build models for groups of facilities with similar temporal water use behavior.
- Ensemble models are most beneficial in facilities with high variance and fewer observations of withdrawal.

1. INTRODUCTION

Sustainable water resources management requires accurate models, predictions, and projections of water demand. Short-term water use forecasting can be crucial in drought management and utility operations. Longer-term projections of water use can help identify potential supply risks under conditions of population growth (Vörösmarty *et al.* 2000) and climate change (Brown *et al.* 2013; Fiorillo *et al.* 2021). These models also form an important component of integrated water systems models and decision support systems that simulate hydrologic water supply, infrastructure, demand, and reuse (Willuweit & O'Sullivan 2013; Sharvelle *et al.* 2017). Accurate models and projections of water demand are especially valuable in locations where water management institutions have relatively limited control on water use. For instance, this is the case in many areas of the Eastern U.S. where large portions of withdrawal are not subject to permitting requirements (Virginia Department of Environmental Quality 2022). However, the factors that govern water demand are highly complex and involve interactions between climatic and environmental conditions, socio-economic factors, pricing, and institutional governance structures. Given this complexity, it is unsurprising that many water use forecasts turn out to be inaccurate in hindsight (Pacific Institute 2013; Perrone *et al.* 2015).

Recognizing this need, numerous studies have used statistical regression models to identify the environmental, socio-economic, and institutional factors associated with greater volumes of water use. For instance, multiple studies have demonstrated the relationship between climatic conditions, land use, and water use at the municipal scale (Balling *et al.* 2008; House-Peters

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

et al. 2010; Mini *et al.* 2014; Lee *et al.* 2015; Toth *et al.* 2018). Several studies have leveraged water use data to characterize drivers of broad-scale geographic variability in per-capita municipal water use efficiency and trends (Sankarasubramanian *et al.* 2017; Worland *et al.* 2018; Chinnasamy *et al.* 2021). Because the factors that influence water use tend to be complex and nonlinear, there is increasing use of machine learning to model and predict water use. Machine learning models have been widely applied in the prediction of physical hydrologic systems (e.g., Akrami *et al.* 2014; Guimarães Santos & Silva 2014; Alizadeh *et al.* 2017a, 2017b). Methods including random forests, boosted regression trees, and artificial neural networks have been leveraged to identify climatic and governance factors that influence municipal and irrigation demand (Toth *et al.* 2018; Bolorinos *et al.* 2020; Fiorillo *et al.* 2021; Lamb *et al.* 2021). Short-term urban demand forecasting has also benefited from methods such as long short-term memory networks (Hu *et al.* 2019; Mu *et al.* 2020; Fu *et al.* 2022; Zanfei *et al.* 2022), neural networks (Huang *et al.* 2021; Huang *et al.* 2022; Liu *et al.* 2023), and hybrid approaches (Guo *et al.* 2022). When compared with linear regression approaches, machine learning models are often able to achieve lower predictive errors than standard approaches (Toth *et al.* 2018; Bolorinos *et al.* 2020; Wongso *et al.* 2020), pointing toward their potential value in water use modeling.

Across this body of research, one factor that is rarely explicitly considered is the impact of data structure on model predictions and inferences. Water use data are inherently hierarchical, with multiple options for grouping and categorizing observations. For instance, water use datasets often include observations through time for multiple water users. These water users in turn can be grouped or classified based on geographic location, water use sector, or institutional governance structures. Depending on their structure, regression approaches may be capturing different drivers of variability that lead to different management implications. For example, models of cross-sectional variability (where there is a single record, such as a long-term average withdrawal, for each water user) and locations can assist in targeting conservation measures (Deoreo & Mayer 2012; Suero *et al.* 2012). Models of temporal variability (where multiple observations through time are available) can lead to more accurate predictions of water use under different policy and drought conditions (Hester & Larson 2016).

These different approaches can provide greater insights into the nature between water use and various factors that influence it. For example, cross-sectional analyses have found a positive correlation between income and water use (Balling *et al.* 2008; House-Peters *et al.* 2010; Sankarasubramanian *et al.* 2017) that is not present in longitudinal studies (Shorridge & DiCarlo 2020). This suggests that water use is greater in locations or households with higher incomes, but not necessarily during periods of greater economic growth. Recognizing this, longitudinal regression has become a standard statistical approach in modeling water use (Polebitski & Palmer 2010; House-Peters & Chang 2011; Baerenklau *et al.* 2014; Shorridge & DiCarlo 2020), where model parameters can vary across groups within population-level constraints. This provides a middle ground between pooled regression models, where all observations are grouped together and described via a single set of model parameters, and unpooled regression where a unique model is fit for each group in the data (Gelman & Hill 2007).

Recent advances in machine learning have begun to develop new approaches that account for hierarchical data structures. For example, the mixed effects random forest (MERF) approach models individual predictions through time as an additive function of a random forest (RF) model of population-level mean behavior processes and individual-level random effects (Hajjem *et al.* 2014; Capitaine *et al.* 2021). Several studies have proposed methods that integrate regression and classification trees within a mixed modeling framework to address subgroups and hierarchies in clinical trial data (Fokkema *et al.* 2018; Seibold *et al.* 2019; Fokkema *et al.* 2021). Other methods leverage ensemble learning, where predictions from multiple models are aggregated into a single prediction (Eygi Erdogan *et al.* 2021). Ensemble learning, in which multiple models are independently fit to a dataset and averaged into a single prediction, has been found to generally reduce model variance which results in more accurate predictions on new data (Kuncheva 2014; James *et al.* 2021). This aspect of ensemble modeling has the potential to improve accuracy in water use prediction, particularly due to previously observed issues with data quality and errors that are present in many water use datasets (Zhang & Balay 2014; Chini & Stillwell 2017; McCarthy *et al.* 2022). Beyond the general approach of leveraging machine learning models within a hierarchical data structure, many of these previous studies also present examples of context-specific algorithm development as they were specifically designed to be compatible with clinical trial data. The development of hierarchical machine learning modeling approaches tailored to water use data has the potential to both increase predictive accuracy relative to current methods and provide new inferences about water use behavior and influences across different hierarchical levels.

The objective of this research was to develop and assess a novel algorithm for prediction of water withdrawals across multiple usage sectors using an ensemble of predictive regression models fit at different hierarchical levels. This work leverages

29 years of monthly withdrawal data from approximately 2,500 water using facilities across Virginia. Models were fit at different grouping levels, ranging from single-facility models to sector-wide models using multiple climatic and socio-economic predictor variables. A cluster analysis was conducted to identify clusters of facilities with similar temporal patterns of water withdrawal and fit cluster-level models. Grouping level models were then combined into a weighted ensemble prediction using quadratic programming. The predictive accuracy of all models was evaluated through a repeated holdout cross-validation approach, and compared to a null model where facility-level withdrawal was based on long-term averages. Finally, the facility-level characteristics associated with improved ensemble predictions were identified to better understand the conditions in which ensemble modeling provides the most value.

2. METHODS

2.1. Data sources and processing

This analysis used long-term records water withdrawal provided by the Virginia Department of Environmental Quality (VDEQ). All water users in the U.S. state of Virginia who withdraw more than 37,854 l (10,000 U.S. gallons) per day are required to report monthly water withdrawal to VDEQ. This dataset includes 313,321 nonzero monthly withdrawal records between 1990 and 2018 from 2,579 water using facilities across eight water use sectors (Table 1). Note that agriculture refers to livestock and agricultural processing operations, rather than crop irrigation. Additional details on withdrawal data are presented in Shortridge & DiCarlo (2020). However, many of these facilities only have short-term records of water withdrawal or a majority of months with zero reported withdrawals. To ensure that all facilities had sufficient data available for model training, weighting, and validation, only facilities with at least 36 nonzero withdrawal observations were retained for inclusion in the analysis. This number was selected because at least 2 years of data are needed to calculate withdrawal anomalies, and at least 2 additional years are needed to split the data into testing and training datasets.

Water withdrawal volumes across different users often vary across several orders of magnitude and exhibit seasonal patterns. In these instances, the use of monthly anomaly values that represent the degree to which a value differs from the long-term average for that month can better account for seasonal variability compared to the use of raw values (Shortridge et al. 2016). To address this variability, all water withdrawal records were converted to water withdrawal anomalies as in Equation (1):

$$W_AN_{f,t} = \frac{W_O_{f,t} - \overline{W_O}_{f,m}}{sd(W_O_{f,m})} \quad (1)$$

where $W_AN_{f,t}$ is the withdrawal anomaly in facility f at time period t ; $W_O_{f,t}$ is the observed withdrawal in facility f at time t , $\overline{W_O}_{f,m}$ is the average withdrawal in facility f for month m ; and $sd(W_O_{f,m})$ is the standard deviation of withdrawal in facility f during month m . Note that estimating anomalies thus requires at least 2 years of data for each facility.

Table 1 | Summary of withdrawal data used in model development

Sector	All data		Retained for analysis		
	Facilities (n)	Observations (n)	Facilities (n)	Observations (n)	Total water use (MG/month)
Agriculture (Ag)	155	7,032	36	5,914	129
Aquaculture (Aq)	14	2,978	12	2,913	866
Commercial (Com)	463	55,573	292	52,721	740
Industrial (Ind)	211	37,464	154	36,968	16,000
Irrigation (Irr)	727	23,036	187	18,157	1,310
Mining (Min)	91	14,394	70	14,260	1,320
Municipal (Mun)	894	166,747	735	164,718	24,900
Thermoelectric (Thm)	24	6,097	23	6,073	201,000
<i>Total</i>	<i>2,579</i>	<i>313,321</i>	<i>1,509</i>	<i>301,724</i>	<i>246,000</i>

A total of 13 socio-economic variables were included as potential predictors of water withdrawal, representing a variety of population, economic, and land-use characteristics that have been shown to have relationships with water withdrawals in previous research (Sankarasubramanian *et al.* 2017; Worland *et al.* 2018; Shortridge & DiCarlo 2020). Additionally, three climatic predictor variables were included to account for widespread evidence of the relationship between weather and water withdrawals (House-Peters & Chang 2011; Brown *et al.* 2013; Lee *et al.* 2015). Additional details on predictor variable data, sources, processing, and formatting are provided in Supplementary material.

2.2. Modeling approach

The predictor variables described above were used to estimate monthly water withdrawal at the facility level using an ensemble of models fit across different grouping levels. Grouping level refers to the specificity of data included to fit the model and included a facility level, sector level, and two cluster-level groupings. A summary and rationale for the different grouping levels included in the analysis is presented in Table 2. At each level, multiple model formulations were tested and the best model in terms of out-of-sample predictive error minimization was retained. These models were combined into a multi-level ensemble model that predicted withdrawal as a weighted average of predictions from the different grouping level models. Model performance was quantified via a repeated cross-validation approach where the data were randomly partitioned at each iteration into distinct training, weighting, and testing datasets. An overview of this process is shown in Figure 1, and additional details are presented in the following sections.

2.2.1. Facility grouping and clustering

The water withdrawal data used in this study can be grouped at multiple hierarchical levels. Each water using facility has multiple observations of water use through time. Facilities are often categorized by water use sector, under the assumption that two facilities in the same water use sector will exhibit similar water use behavior. For this study, predictive models were fit at four different levels of facility grouping: facility level, sectoral level, small cluster level, and large cluster level (Table 2). At the finest level, facility-level grouping entailed fitting a distinct model for each facility in the dataset. This allows for the model to be highly tailored to the water use characteristics of that facility but less generalizable to new data, particularly in instances where a facility does not have many observations to draw from (Gelman & Hill 2007). The next level of grouping was the sectoral level, where a single model was fit to all facilities within that sector. This provides a representation of generalized water use patterns in a given sector, such as the higher irrigation withdrawals that are observed during periods of high temperature and low rainfall (Shortridge & DiCarlo 2020). This provides a model of how sectoral water withdrawals relate in general with predictor variables but will likely result in less accurate predictions for a single facility.

One limitation with sectoral grouping is that facilities in a single sector might actually exhibit very different patterns of water use (Attaallah 2018; McCarthy *et al.* 2022). Thus, the small and large cluster grouping levels were determined based

Table 2 | Summary of models included in comparison and rationale for inclusion

Model name	Description and rationale
Facility-grouping level	Separate model fit to each facility in the dataset. Captures facility-level water use behavior, but not generalizable to other facilities.
Sector-grouping level	Model fit using data from all facilities within each water use sector. Captures general water use behavior across multiple facilities at the expense of accuracy at individual facility level.
Large cluster grouping level	Model fit using data from all facilities within each large cluster ($k = 3$). Clusters are defined based on temporal water use patterns, and thus contain facilities with similar withdrawal patterns even if they are different water use sectors.
Small cluster grouping level	Model fit using data from all facilities within each small cluster ($k = 8$). Same as large clusters, but with facilities partitioned into smaller groups with less in-group variability in temporal withdrawal patterns.
Ensemble	Withdrawal predictions are a weighted average of the four grouping level models above.
Null	Withdrawal predictions are equal to the long-term average withdrawal in each month for each facility. Included as a baseline for comparison.

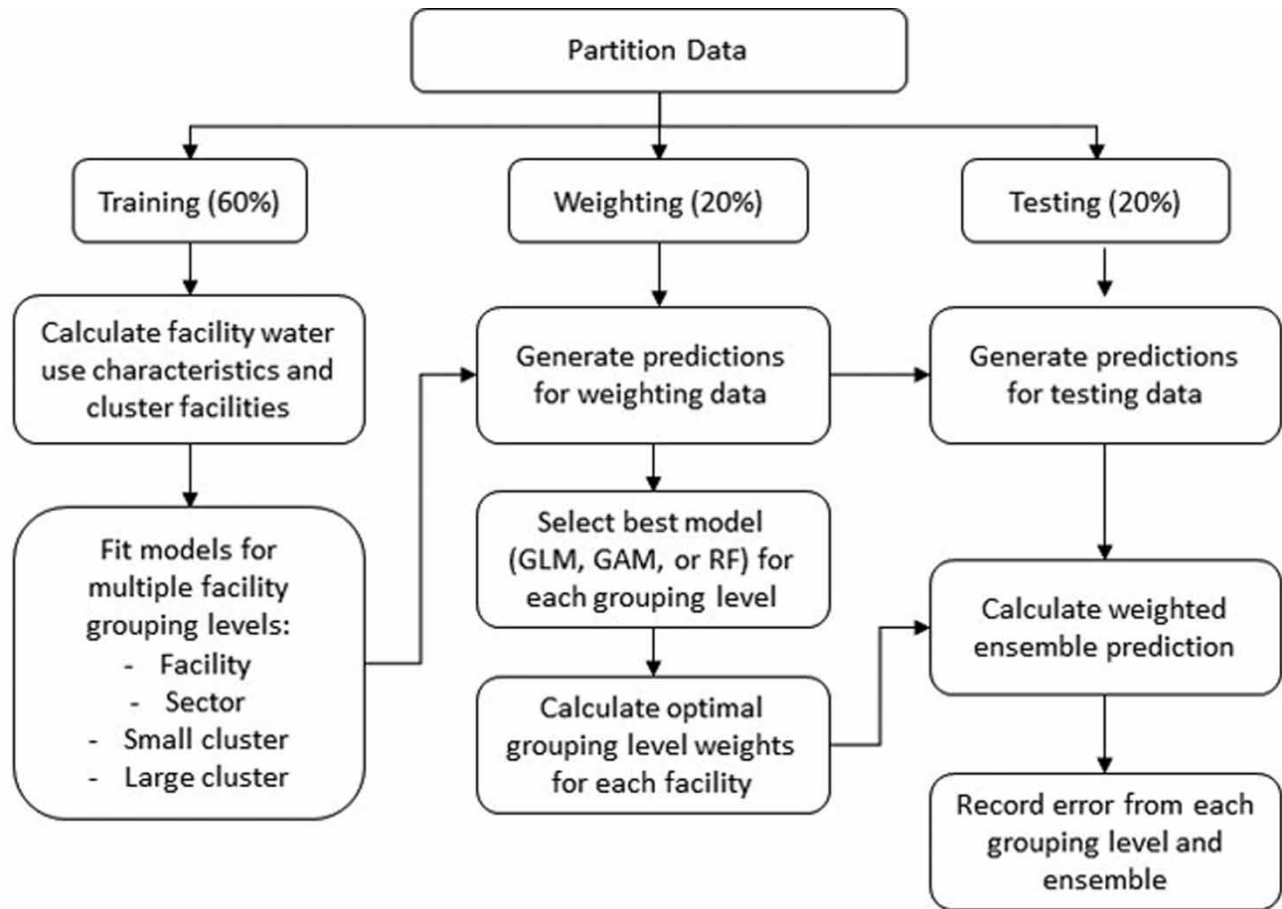


Figure 1 | Diagram overview of modeling approach used during each iteration of cross-validation.

on the results of a hierarchical cluster analysis (Everitt *et al.* 2011) that identified coherent facility groupings based on five water use characteristics calculated for each facility:

- Mean withdrawal volume (MG/month), log transformed.
- Coefficient of variation: standard deviation of withdrawal divided by mean.
- Seasonality: the lowest 3-month mean withdrawal divided by the highest 3-month mean withdrawal, where lower values indicate greater seasonality in withdrawal volume.
- Autocorrelation: maximum degree of autocorrelation observed at any time lag.
- Number of observations: the number of nonzero withdrawal observations available.

To determine the optimal number of clusters, facilities were divided into $k \in \{1, 15\}$ hierarchical clusters based on Euclidean distance. Gap statistic estimates for each value of k exhibited non-monotonic behavior indicating well defined clusters at $k = 3$ and $k = 8$, suggesting that there were three large clusters of facilities that could be further divided into eight smaller subclusters (Tibshirani *et al.* 2001). An analysis of correspondence between cluster assignment and sector indicated that clusters generally did not correspond to a single sector. This suggests that there are certain patterns of water use behavior that cannot be explained simply by sectoral classifications, consistent with previous research (Attaallah 2018; McCarthy *et al.* 2022). Thus, models were also fit at the large ($k = 3$) and small ($k = 8$) cluster levels, where data from all facilities within a single cluster were combined into a single model. Additional details and results of the cluster analysis are included in Supplementary material.

2.2.2. Classical regression and machine learning models

For each of the grouping levels described above, multiple classical regression and non-parametric machine learning approaches were compared to identify the most effective predictor of water withdrawals. The general formulation used in modeling withdrawal anomalies is shown in Equation (2), where $\mathbf{W_AN}_f$ is a vector of anomaly withdrawal predictions of length t in facility f , where t is the number of months of observations available in the training dataset. These were estimated using a generalized function of a $m \times t$ matrix of m predictor variables across t months \mathbf{X}_f , plus an error term ε .

$$\mathbf{W_AN}_f = f(\mathbf{X}_f) + \varepsilon \quad (2)$$

Three different forms for the functional relationship f were tested at each grouping level. The first was a Gaussian linear regression (GLM) model. The second was a semi-parametric Gaussian generalized additive model (GAM), where smoothing functions are applied to the predictor variables to capture non-linear relationships between predictor and response variables without *a priori* assumptions about the form of that relationship (Hastie & Tibshirani 1986). GAM models were fit using the mgcv package in R (Wood 2011). The final model form was a nonparametric RF model, where predictions from multiple rule-based regression trees are combined into a single prediction (Breiman 2001). RF models were fit using the randomForest package in R (Liaw & Wiener 2002). All model predictions were then converted from anomaly values back to a vector of withdrawal predictions as in Equation (3) prior to estimating model error:

$$\mathbf{W_P}_f = \mathbf{W_AN}_f \times sd(\mathbf{W_O}_{f,m}) + \overline{\mathbf{W_O}}_{f,m} \quad (3)$$

For each grouping level model, the GLM, GAM, and RF models were compared in terms of their mean absolute error (MAE) across the weighting dataset, with the lowest error model retained. This identified the formulation with the lowest out-of-sample predictive error. Following this process, each facility had four sets of withdrawal predictions for the weighting dataset generated by models fit at the facility, sector, small cluster, and large cluster grouping level. These predictions were then combined into a weighted ensemble prediction as follows:

$$\mathbf{W_P_Ens}_f = \mathbf{w}_f \cdot \mathbf{W_P}_{f_all} \quad (4)$$

where \mathbf{w}_f is a facility-specific vector of weights summing to 1.0, and $\mathbf{W_P}_{f_all}$ is an $n \times 4$ matrix of predictions from the four different grouping level models (facility, sector, large cluster, and small cluster) for the n weighting data observations for facility f . The resulting $\mathbf{W_P_Ens}_f$ is thus a vector of n predictions obtained from a weighted average of individual grouping level model predictions, with the weights optimized to minimize error for the weighting dataset. The values of the weights \mathbf{w}_f for each facility were estimated using a quadratic programming problem (Goldfarb & Idnani 1983) implemented via the quadprog package in R (Turlach *et al.* 2019) of the form:

$$\text{Minimize } (\mathbf{W_O}_f - \mathbf{w}_f \cdot \mathbf{W_P}_{f_all})^2 \quad (5)$$

Subject to

$$\sum_{i=1}^4 w_i = 1.0$$

$$w_i \geq 0.0 \quad \forall i \in [1, 4] \quad (6)$$

The four grouping level models, as well as the ensemble model, were then used to generate withdrawal predictions for the testing dataset. Thus, the testing dataset was not used in either the initial model fitting or the ensemble weighting.

2.3. Model evaluation

To evaluate models in terms of their out-of-sample predictive accuracy, a 100-fold cross-validation approach was used (Hastie *et al.* 2009). At each iteration, the data were partitioned into three groups, with approximately 60% assigned to model training, 20% to model weighting, and 20% to model testing. This partitioning was done based on years, so that in each holdout

iteration the years of data assigned to model training, weighting and testing were consistent across facilities, sectors, and clusters. This also ensured that a single year of withdrawal data were not split into different partitions. The training data were used to fit three models of different functional forms (GLM, GAM, and RF) for each grouping level. These models were then used to predict withdrawal in the weighting dataset, with a single functional form selected for each grouping level based on root mean absolute error (RMAE). The predictions from the selected models were then used to determine the weights used in the ensemble model. Finally, the grouping level and ensemble models were used to predict withdrawals in the testing dataset. To provide a baseline for evaluating model performance, these predictions were compared to a null model where each prediction of monthly water use was equal to the long-term monthly mean value for that facility. Thus, the null model captured seasonal variation for each facility but did not include climatic or socio-economic factors that could induce variation beyond seasonal patterns. This approach of using long-term average withdrawal is often necessary in water supply planning contexts when more complex models are unavailable, and thus provides a reasonable baseline for comparison with other model formulations. MAE across the testing dataset was calculated for each model and facility by averaging absolute differences between observed and predicted withdrawal in each observation n across each holdout iteration h (Equation (8)). Because absolute errors tend to scale with withdrawal volume, relative mean absolute error (RMAE), where MAE was presented as a fraction of mean facility withdrawal, was also calculated to allow for comparison of error across facilities with different magnitudes of withdrawal (Equation (9)).

$$\text{MAE}_f = \frac{1}{H} \sum_{h=1}^H \frac{1}{N} \sum_{n=1}^N \text{abs}(W_{-O_{f,n}} - W_{-P_{f,n}}) \quad (7)$$

$$\text{RMAE}_f = \frac{1}{H} \sum_{h=1}^H \frac{\text{MAE}_{f,h}}{W_{-O_{f,h}}} \quad (8)$$

To better understand the characteristics associated with facilities where ensemble modeling provided the most benefit relative to facility-specific models, the difference between RMAE from the facility grouping and ensemble models (Equation (10)) was linearly regressed against facility water use characteristics C . This included the logarithm of mean facility withdrawal, the number of nonzero observations, the coefficient of variation, autocorrelation, seasonality, and a use sector, as defined in Section 2.2.1.

$$\text{EnsImp}_f = \text{RMAE}_{f_{-fac}} - \text{RMAE}_{f_{-ens}} \quad (9)$$

$$\text{EnsImp}_f = \alpha + \beta C_f + \varepsilon \quad (10)$$

3. RESULTS

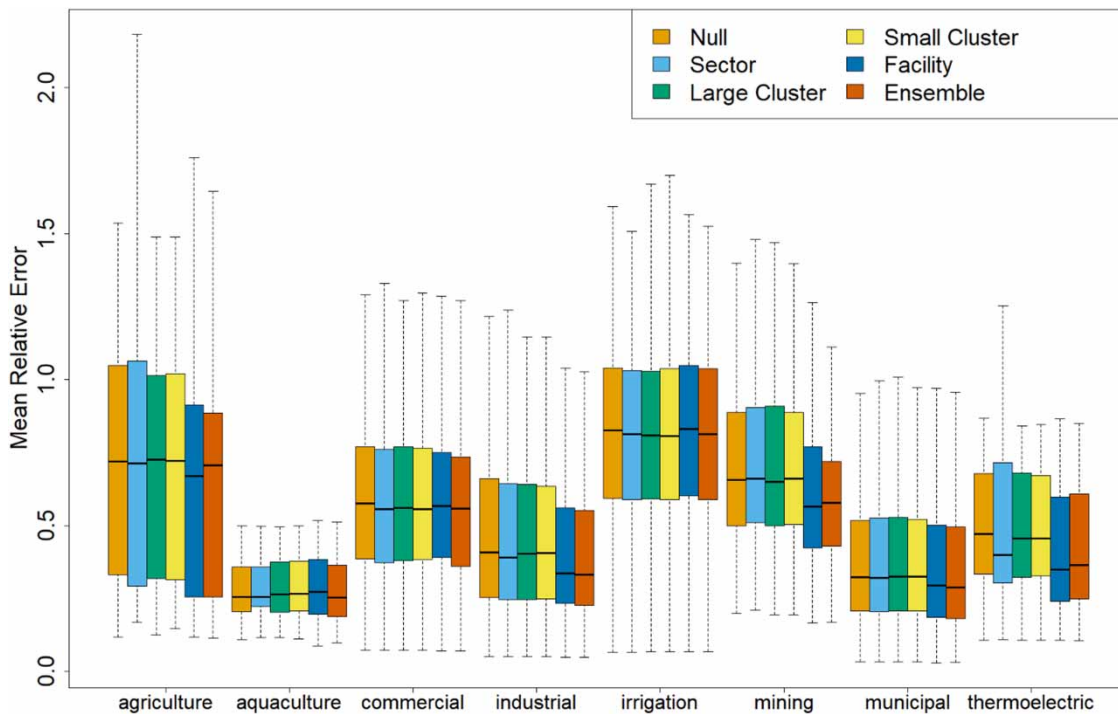
3.1. Model selection and performance

In each iteration of the holdout cross-validation, three model formulations (GLM, GAM, or RF) were compared for each of the four grouping levels (sector, large cluster, small cluster, and facility) based on out-of-sample RMAE in the weighting dataset. The formulation with the lowest RMAE was selected for the grouping level model and for incorporation into the ensemble model in that holdout iteration. The frequency with which each formulation was selected (i.e., minimized out-of-sample RMAE) at each grouping level is presented in Table 3. For the cluster-level grouping and for most sector-level groupings, the most frequently selected models were the GLM formulation. More complex formulations (GAM and RF) were more often selected for the facility-level grouping models and for the agriculture, mining, and municipal. The relatively strong performance of the simpler linear models could be due to a potential for overfitting with the GAM and RF formulations, where their flexibility results in a lower bias relative to model training data but greater variance and error when fit to new datasets (Hastie *et al.* 2009).

Boxplots summarizing the distribution of RMAE for testing dataset predictions across all 100 holdout cross-validations is presented in Figure 2. For all model structures, relative error depended strongly on the sector assessed, with the highest predictive errors in agricultural and irrigation sectors. Variations in error across model structures were relatively small compared to differences across sectors and the variation in error across facilities within a sector (represented by the size of each box).

Table 3 | Frequency at which each model formulation achieved the lowest out-of-sample error across different grouping levels

	GLM	GAM	RF
Sector-level groupings			
Agriculture	47.5%	19.2%	33.3%
Aquaculture	62.6%	9.1%	28.3%
Commercial	53.5%	41.4%	5.1%
Industrial	52.5%	41.4%	6.1%
Irrigation	61.6%	33.3%	5.1%
Mining	31.3%	45.5%	23.2%
Municipal	24.2%	35.4%	40.4%
Thermoelectric	51.5%	18.2%	30.3%
Cluster-level groupings			
Large cluster	55.2%	22.2%	22.6%
Small cluster	46.5%	25.4%	28.2%
Facility-level groupings			
Facility	28.6%	16.7%	54.4%

**Figure 2** | Relative mean absolute errors across facilities in each water use sector. Boxplots show the distribution of RMAE for all facilities in each sector averaged across cross-validation iterations.

Because null model performance, based only on long-term monthly average withdrawal for each facility, also varied significantly between sectors assessed, this variation in performance across sectors is likely due to differences in the typical variance of withdrawals through time. The sectors where RMAE were the lowest (aquaculture, industrial, municipal, and thermoelectric) are also those where the coefficient of variation in withdrawal tend to be the lowest (see Supplementary material, Figure S2). In sectors where withdrawals exhibit more variability through time (agriculture and irrigation), demonstrated

by higher coefficient of variation values, predictive RMAE tended to be higher for all model forms assessed. This is unsurprising as agricultural and irrigation water use can change substantially through time as growers rotate crops, changing production practices, or introduce new irrigation management technologies.

A summary of mean RMAE for each model grouping level is presented in Table 4. The ensemble model had the lowest mean RMAE in all sectors except agriculture and industrial, where the facility-grouping level models had the lowest mean RMAE. Paired, two-sided Wilcoxon rank sum tests were used to compare the distribution of RMAE in the ensemble and facility-grouping models for each sector. In the aquaculture, commercial, irrigation, mining, and municipal sectors, the use of the ensemble formulation resulted in statistically significant reductions in RMAE compared to the facility-level models. The only sector in which the ensemble model resulted in a statistically significant increase in error was agriculture.

Figure 3 presents a summary of the percentage of facilities where predictions are improved through the use of a model ensemble relative to the other model forms. Across all facilities, the use of a model ensemble results in lower errors in 63–65% of facilities, depending on the model form with which it's compared. The ensemble process results in the broadest improvement for all non-facility models in the thermoelectric sector. For thermoelectric facilities, the ensemble model resulted in lower errors in anywhere from 78.3 to 95.7% of facilities when compared to the null, sector, and cluster grouping

Table 4 | Mean RMAE across all holdout iterations for different water use sectors and grouping level formulations

Sector	Null	Sector	Large cluster	Small cluster	Facility	Ensemble	p-value
Agriculture	1.623	1.471	1.595	1.513	1.364	1.377	7.54×10^{-5}
Aquaculture	0.333	0.330	0.340	0.341	0.316	0.313	5.00×10^{-3}
Commercial	0.667	0.658	0.658	0.659	0.651	0.638	$<10^{-5}$
Industrial	0.560	0.576	0.571	0.572	0.438	0.460	5.37×10^{-1}
Irrigation	0.879	0.862	0.870	0.869	0.851	0.846	$<10^{-5}$
Mining	0.822	0.835	0.831	0.831	0.774	0.738	$<10^{-5}$
Municipal	0.601	0.572	0.578	0.572	0.538	0.508	$<10^{-5}$
Thermoelectric	0.533	0.526	0.528	0.530	0.472	0.471	9.31×10^{-1}

Bold values indicate the grouping level with the lowest mean RMAE for that sector. p-values refer to the significance level of a two-sided paired Wilcoxon Rank Sum test between the facility grouping and ensemble models.

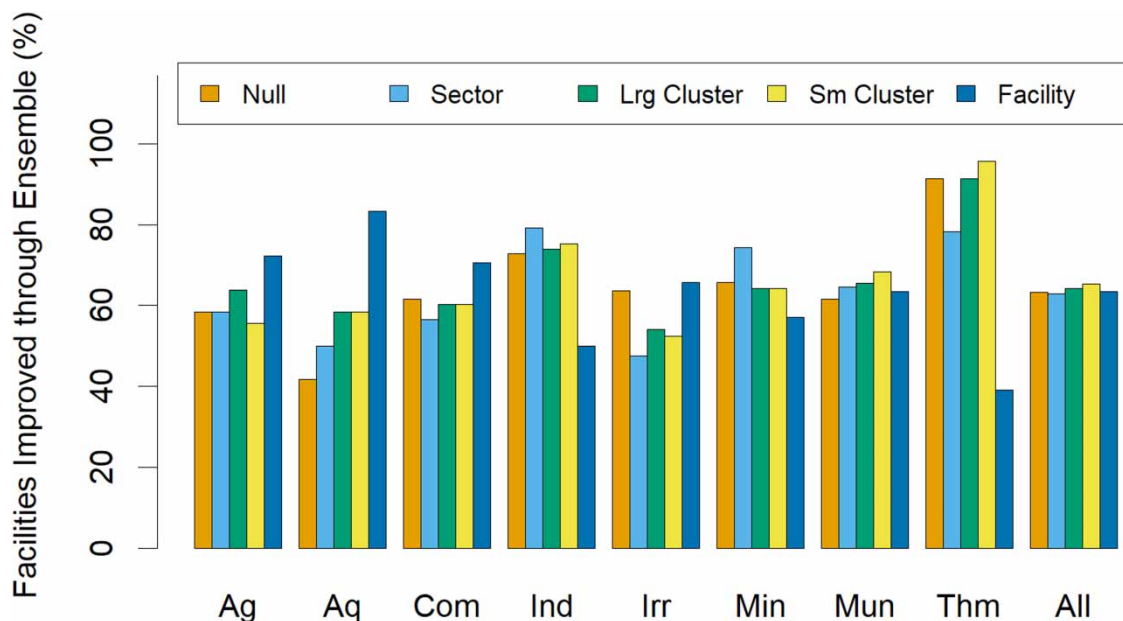


Figure 3 | Percentage of facilities where ensemble model reduced errors relative to null and grouping level models.

models. However, it only resulted in improved performance in 39.1% of thermoelectric facilities when compared to facility-specific models. Similar behavior is observed for industrial facilities. Ensemble improvements were most consistent in the municipal sectors, where errors were reduced by 61.5–68.3% of facilities, regardless of the grouping level model with which the ensemble is compared. These results demonstrate that the ensemble process is likely to result in better predictions relative to single grouping level models when applied to most facilities. In this sense, its value is in providing a general approach that could be applied to many facilities across a broad, heterogenous dataset, rather than being an optimum approach for a single facility.

3.2. Ensemble model structure

To better understand the relative influence that different individual grouping level models play within the ensemble predictions, Figure 4 shows density plots of the average weights for each grouping level model for all facilities in each sector. In all sectors, facility-level weights were generally higher than weights for other grouping levels models. It is notable that the sector-level weights were generally no higher than the cluster weights, and in several sectors (agriculture, aquaculture, and thermoelectric), the small cluster weights were often higher than sector-level weights. This is potentially due to variation across facility-level water use characteristics in these sectors, meaning that a sector-level model is less capable of accurately predicting water use in an individual facility. For example, the agricultural withdrawal data are unique in that it includes some facilities with highly seasonal water withdrawals, and other facilities with only minor seasonality (see Supplementary material, Figure S4). These are also the three sectors with the smallest number of facilities (Table 1), meaning that they provide an insufficient sample size to derive generalizable, sector-level withdrawal behavior. This demonstrates how other facilities with similar water use patterns, even if they are in a different sector, can be leveraged to provide more accurate predictions of water withdrawal.

To better understand the facility characteristics associated with improved ensemble model performance relative to facility-grouping models, the predictive improvement from use of a model ensemble for each facility was regressed against facility water use characteristics. The results of this regression are presented in Table 5. The ensemble model tended to provide the most improvement relative to the facility-grouping model in facilities with a lower number of observations, higher coefficient of variation, and less autocorrelation. These are all conditions that create a potential for facility model training datasets that are less representative of withdrawal behavior as a whole due to smaller sample size and greater data variance through time. Because this can result in models that are overfit to training data and less generalizable to unseen data, the incorporation of other, more general model formulations into an ensemble can provide particularly noticeable reduction in out-of-sample predictive errors in this context.

4. DISCUSSION

The use of ensemble modeling resulted in a statistically significant reduction in mean out-of-sample RMAE in five of the eight sectors assessed (aquaculture, commercial, irrigation, mining, and municipal) relative to models fit using just facility-specific data. The only sector in which the ensemble model resulted in a statistically significant increase in error was agriculture. Across all facilities, the ensemble model resulted in error reduction relative to grouping level models in over 60% of facilities. Thus, while it does not necessarily result in greater predictive accuracy for all facilities in this dataset, it does result in predictive improvements across the population of water users as a whole. In this sense, its value is likely highest in situations where a general modeling approach is needed to simulate longitudinal water withdrawals across a heterogenous body of water users, rather than a model of a single water user, especially for users that have a long and robust record of water withdrawals. Particularly for large water users (especially thermoelectric facilities) that tend to dominate overall withdrawal volumes (McCarthy *et al.* 2022), facility-specific models will likely prove most accurate. However, our results demonstrate that there are numerous other types of water use where ensemble modeling provides predictive value.

The regression results in Table 5 indicate that the improvement in using a model ensemble is particularly large in facilities with fewer observations, greater variance, and less autocorrelation. These are all conditions where there is a potential for greater discrepancy between model training and testing datasets, likely resulting in models that are less generalizable to unseen data. This result mirrors previous discussions of pooled and unpooled regression models, where unpooled models fit tend to be overfit and less generalizable when a small number of observations are available (Gelman & Hill 2007). Figure 5 shows an example of this behavior across a single cross-validation of two facilities. The top panel of the figure shows results for a municipal facility where the facility-level model outperformed the ensemble model in terms of predictive error in the

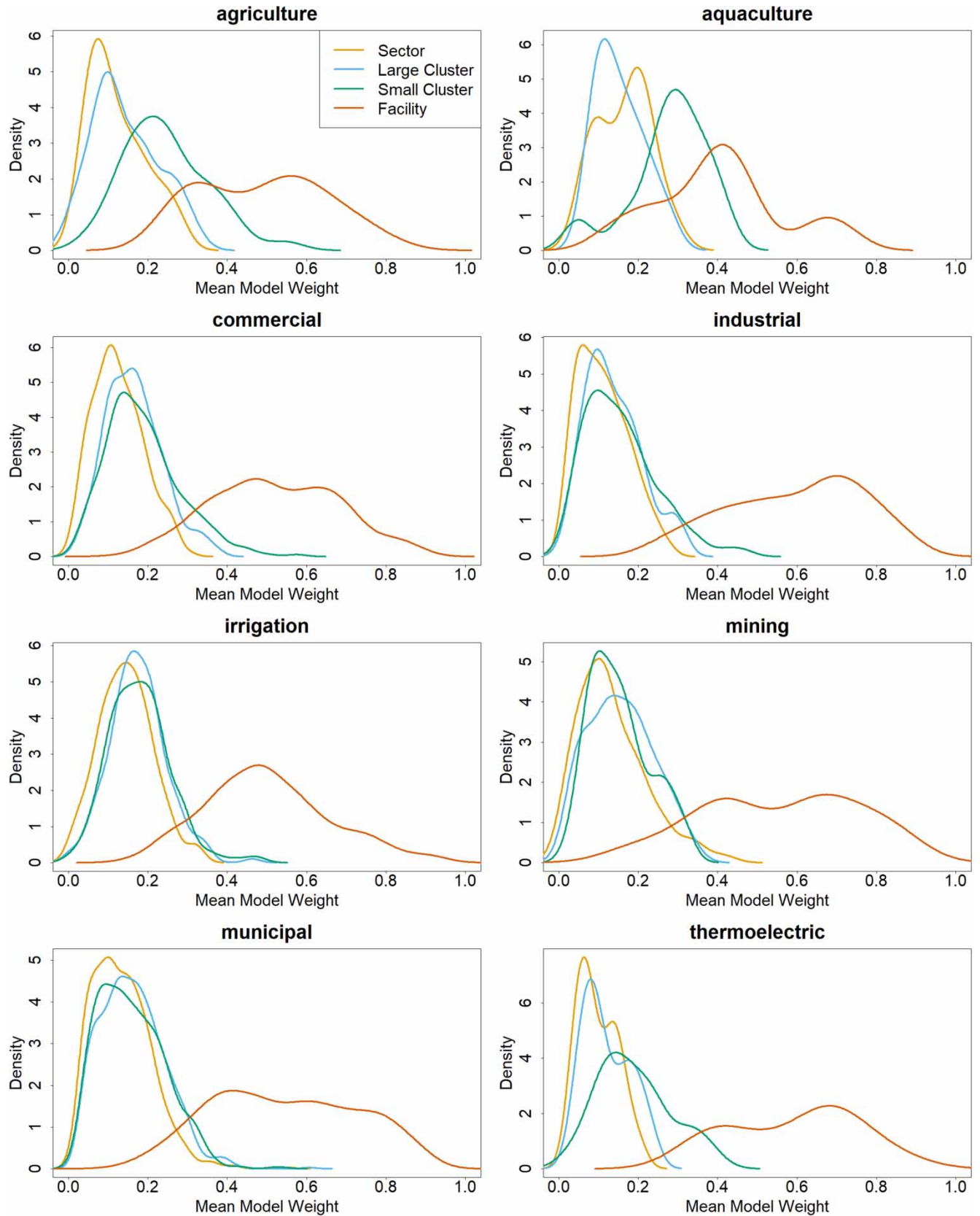
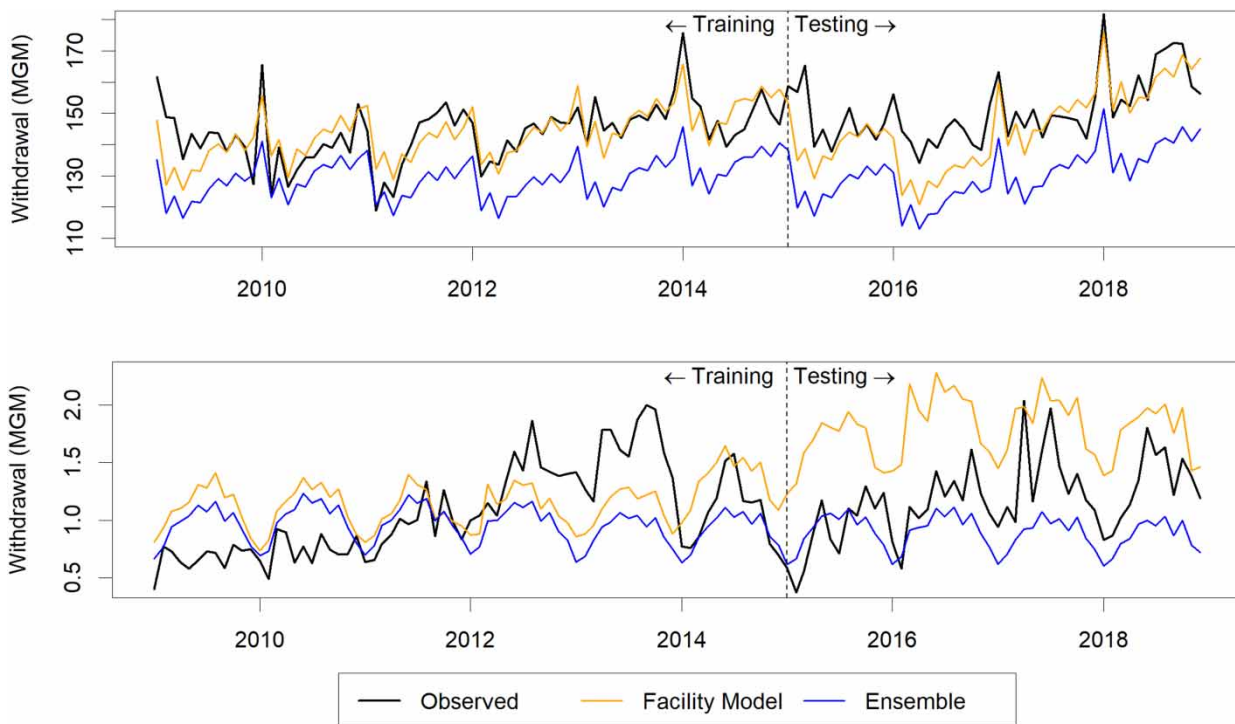


Figure 4 | Density plot of mean ensemble model weights across facilities in each sector, averaged across cross-validation iterations.

Table 5 | Factors associated with improved ensemble performance relative to facility model performance

	Estimate	Std. Error	p-value
Intercept	1.40×10^{-02}	8.00×10^{-05}	7.80×10^{-02}
log(Water.Use.MGM)	-1.00×10^{-03}	1.00×10^{-05}	1.27×10^{-01}
n.obs.nonzero	-2.85×10^{-05}	1.13×10^{-05}	1.20×10^{-02}
Water.Use.COV	1.60×10^{-02}	1.00×10^{-05}	< 0.001
Water.Use.ACF.strength	-2.20×10^{-02}	4.00×10^{-05}	< 0.001
Water.Use.Seasonality	$7.00E \times 10^{-05}$	5.00×10^{-05}	1.86×10^{-01}
UseType (aquaculture)	3.00×10^{-05}	1.30×10^{-02}	8.39×10^{-01}
UseType (commercial)	-1.00×10^{-05}	7.00×10^{-05}	8.32×10^{-01}
UseType (industrial)	-5.00×10^{-05}	8.00×10^{-05}	4.66×10^{-01}
UseType (irrigation)	-1.60×10^{-02}	7.00×10^{-05}	2.60×10^{-02}
UseType (mining)	1.50×10^{-02}	8.00×10^{-05}	7.80×10^{-02}
UseType (municipal)	4.00×10^{-05}	7.00×10^{-05}	6.03×10^{-01}
UseType (thermoelectric)	-5.00×10^{-05}	1.10×10^{-02}	6.31×10^{-01}

**Figure 5** | Example plots from a single cross-validation iteration of (top) a well-fit facility model that accurately predicts withdrawal in the testing time period and results in less error than the ensemble model prediction, and (bottom) an overfit facility model that overestimates a positive trend through time and results in greater error in the testing time period relative to the ensemble model.

testing dataset. In this instance, the facility-level model was well-fit to the training data but still generalizable to the testing data. The bottom panel shows results for an industrial facility where the facility-level model was overfit to the training data; in particular, it overestimated a positive trend through time that did not extend into the testing period. In this case, including other level models in an ensemble prediction resulted in lower error in the testing dataset. High variance in water use observations could also potentially be indicative of previously observed data quality issues in self-reported water

use (Zhang & Balay 2014; Chini & Stillwell 2017; McCarthy *et al.* 2022). Because individual models may be highly sensitive to outlier or erroneous data, the ability to reduce the impact of these errors in predictive models is another potential benefit to the ensemble modeling approach.

These results are consistent with other studies that have found a frequent benefit of using machine learning approaches to predict water withdrawal when compared to linear models (e.g., Toth *et al.* 2018; Bolorinos *et al.* 2020; Wongso *et al.* 2020). In this study's model selection process, non-linear GAM and RF models were selected most often for agriculture, mining, and municipal sector-level models, as well as for the majority of facility-level models (Table 3). In these instances, the added flexibility of non-linear approaches seems to provide the most value. However, the other sector-level models and the large cluster models most often used GLMs, indicating that in these instances a simple linear approach is suitable. This suggests a potential for overfitting when using machine learning approaches (Hastie *et al.* 2009). Because ensemble learning has been found to generally reduce model variance and overfitting (Kuncheva 2014; James *et al.* 2021), this possibly explains some of the benefit demonstrated by ensemble learning in this study.

These results have several practical implications for water supply management. One notable finding is the relatively low performance of sector-level models. Comparing mean RMAE across all holdout iterations, sector-level models were never the lowest error formulation and in the industrial and mining sectors actually resulted in greater error than a null model based only on long-term average withdrawal alone (Table 4). This suggests that the relationships between socio-economic and climatic conditions with water withdrawal vary too significantly across facilities in those sectors to be beneficial in making facility-level predictions. The better performance of cluster-level models in these sectors, combined with the higher weights attributed to cluster-level models within ensembles (Figure 4), suggest the importance of classifying water users based on usage behavior rather than sectoral classifications alone. It is also important to note that a large body of research on water use focuses on large municipal utilities with long-term records. While the importance and influence of these water users is clear, these results demonstrate that some modeling approaches that are effective in these contexts will be less so among smaller or more newly established water users with shorter and more variable withdrawal records. Alternative methods such as the ensemble modeling approach presented here could be particularly valuable in locations where water use is fairly decentralized and for estimating the impact of newly established water withdrawals.

While these results demonstrate a value in the use of ensemble modeling in predicting withdrawal across a broad, heterogeneous body of water users, there are several limitations that could be addressed through further research. This work is based on data from a single state, and additional research would be needed to demonstrate if the findings here are generalizable to other locations with different climates and regulatory contexts for water use. Additionally, across the 28 years of data included in our dataset, it is possible that certain institutional changes or conditions occurred that would influence water use. For example, more widespread use of water-efficient appliances has led to a documented decrease in household water use (Deoreo & Mayer 2012) and severe drought combined with public educational campaigns lead to reduced outdoor water use (Bolorinos *et al.* 2020). While these unaccounted for variables would not impact the overall conclusions about the potential value of ensemble modeling, they could potentially improve model performance across many formulations.

Several areas of additional research could be envisioned to build on the results presented here. For instance, this research grouped facilities based on temporal water usage characteristics and sectoral classifications. However, water withdrawals could also depend on regulatory governance or water source, as well as geographic location or climate regions. Exploration of alternative grouping strategies could be a valuable area of additional research. Different methods for model evaluation could also advance the results presented here. For instance, this work used a global error metric (RMAE) across all observations, but event detection metrics (Liemohn *et al.* 2021) that quantify the degree to which models capture specific conditions of interest, such as periods of high withdrawal, may be of value, as could deviation-based metrics (Barati *et al.* 2014). Quantifying model outcomes across multiple performance criteria, as in Adnan *et al.* (2023), could also provide a more comprehensive view of model performance. Similarly, this work employs a random cross-validation approach to identify generalizable relationships between withdrawal and predictor variables that are valid across the full period of data. However, practical forecasting needs might be better served by models with sequential training, weighting, and testing periods. While this research compared the ensemble machine learning approach to modeling forms that are commonly applied to withdrawal data, additional research that compares the approach to other hierarchical machine learning methods could lead to further improvements. Finally, the occurrence of unaccounted for times of uncertainty, such as pandemics or natural disasters, will likely lead to conditions and water use behavior that exceeds the range of conditions included in this study but present crucial times for ensuring reliable water supply.

Additional research addressing this question will be critical in improved water management during times of extreme stress.

5. CONCLUSION

Water withdrawal data are inherently hierarchical, often composed of multiple observations for each water user through time, and multiple ways of grouping and categorizing those users. Machine learning models are becoming more widely used in water use modeling, but rarely account for the hierarchical nature of water withdrawal data or make use of this structure to improve predictions. This work presents a novel approach for prediction of longitudinal water withdrawals across multiple usage sectors using an ensemble of machine learning models fit at different hierarchical grouping levels. These grouping levels included facility and sectoral-level models, as well as facility clusters determined based on temporal water use characteristics. Grouping level models were also combined into an ensemble model that predicted withdrawal as a weighted average of predictions from each individual grouping level model. For all model structures, relative error depended strongly on the sector assessed, with the highest predictive errors in agricultural, irrigation, and municipal sectors. The ensemble models achieved statistically significant reductions in error compared to facility-level models in the majority of water use sectors assessed. The use of an ensemble model resulted in more accurate predictions relative to the facility model in 63% of facilities, and ensemble improvements were greatest for facilities with relatively few records and high variance in withdrawal. This points to their potential value in predicting withdrawal for facilities with relatively short records of withdrawal or data quality issues that could lead to highly variable withdrawal estimates. Inspection of the weights used in the ensemble model indicated that small cluster weights were often higher than sector-level weights, pointing toward the limitations of sectoral-level models and the potential benefits of considering the behavior of facilities with similar water use patterns, even if they are in a different sector. The ensemble modeling method presented here can thus provide a general approach for prediction of water withdrawals that can be applied across heterogeneous, multi-sector groupings of water users.

ACKNOWLEDGEMENTS

I would like to gratefully acknowledge the Virginia Department of Environmental Quality for providing the data used in this project.

DATA AVAILABILITY STATEMENT

All code and data used in this analysis are available at: https://osf.io/5pqvx/?view_only=a9b67a7867eb411585897076bf36a433.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

- Adnan, R. M., Sadeghifar, T., Alizamir, M., Azad, M. T., Makarynsky, O., Kisi, O., Barati, R. & Ahmed, K. O. 2023 Short-term probabilistic prediction of significant wave height using Bayesian model averaging: case study of Chabahar Port, Iran. *Ocean Engineering* **272**, 113887. <https://doi.org/10.1016/j.oceaneng.2023.113887>.
- Akrami, S. A., Nourani, V. & Hakim, S. J. S. 2014 Development of nonlinear model based on wavelet-ANFIS for rainfall forecasting at Klang Gates Dam. *Water Resources Management* **28** (10), 2999–3018. <https://doi.org/10.1007/s11269-014-0651-x>.
- Alizadeh, M. J., Kavianpour, M. R., Kisi, O. & Nourani, V. 2017a A new approach for simulating and forecasting the rainfall-runoff process within the next two months. *Journal of Hydrology* **548**, 588–597. <https://doi.org/10.1016/j.jhydrol.2017.03.032>.
- Alizadeh, M. J., Shahheydari, H., Kavianpour, M. R., Shamloo, H. & Barati, R. 2017b Prediction of longitudinal dispersion coefficient in natural rivers using a cluster-based Bayesian network. *Environmental Earth Sciences* **76** (2), 86. <https://doi.org/10.1007/s12665-016-6379-6>.
- Attaallah, N. A. M. 2018 *Demand Dissagregation for Non-Residential Water Users in the City of Logan, Utah, USA*. M.S. Thesis, Civil and Environmental Engineering. Utah State University.
- Baerenklau, K. A., Schwabe, K. A. & Dinar, A. 2014 The residential water demand effect of increasing block rate water budgets. *Land Economics* **90** (4), 683–699. <https://doi.org/10.3368/le.90.4.683>.
- Balling, R. C., Gober, P. & Jones, N. 2008 Sensitivity of residential water consumption to variations in climate: an intraurban analysis of Phoenix, Arizona. *Water Resources Research* **44** (10). <https://doi.org/10.1029/2007WR006722>.

- Barati, R., Neyshabouri, S. A. A. S. & Ahmadi, G. 2014 Development of empirical models with high accuracy for estimation of drag coefficient of flow around a smooth sphere: an evolutionary approach. *Powder Technology* **257**, 11–19. <https://doi.org/10.1016/j.powtec.2014.02.045>.
- Bolorinos, J., Ajami, N. K. & Rajagopal, R. 2020 Consumption change detection for urban planning: monitoring and segmenting water customers during drought. *Water Resources Research* **56** (3). <https://doi.org/10.1029/2019WR025812>.
- Breiman, L. 2001 Random forests. *Machine Learning* **45** (1), 5–32.
- Brown, T. C., Foti, R. & Ramirez, J. A. 2013 Projected freshwater withdrawals in the United States under a changing climate. *Water Resources Research* **49** (3), 1259–1276. <https://doi.org/10.1002/wrcr.20076>.
- Capitaine, L., Genuer, R. & Thiébaud, R. 2021 Random forests for high-dimensional longitudinal data. *Statistical Methods in Medical Research* **30** (1), 166–184. <https://doi.org/10.1177/0962280220946080>.
- Chini, C. M. & Stillwell, A. S. 2017 Where are all the data? The case for a comprehensive water and wastewater utility database. *Journal of Water Resources Planning and Management* **143** (3), 01816005. [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0000739](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000739).
- Chinnasamy, C. V., Arabi, M., Sharvelle, S., Warziniack, T., Furth, C. D. & Dozier, A. 2021 Characterization of municipal water uses in the contiguous United States. *Water Resources Research* **57** (6). <https://doi.org/10.1029/2020WR028627>.
- Deoreo, W. B. & Mayer, P. W. 2012 Insights into declining single-family residential water demands. *Journal – American Water Works Association* **104** (6), E383–E394. <https://doi.org/10.5942/jawwa.2012.104.0080>.
- Everitt, B. S., Leese, M., Stahl, D. & Landau, S. 2011 *Cluster Analysis*. John Wiley & Sons, London, United Kingdom.
- Eygi Erdogan, B., Özögür-Akyüz, S. & Karadayı Atas, P. 2021 A novel approach for panel data: an ensemble of weighted functional margin SVM models. *Information Sciences* **557**, 373–381. <https://doi.org/10.1016/j.ins.2019.02.045>.
- Fiorillo, D., Kapelan, Z., Xenochristou, M., De Paola, F. & Giugni, M. 2021 Assessing the impact of climate change on future water demand using weather data. *Water Resources Management* **35** (5), 1449–1462. <https://doi.org/10.1007/s11269-021-02789-4>.
- Fokkema, M., Edbrooke-Childs, J. & Wolpert, M. 2018 Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods* **50** (5), 2016–2034. <https://doi.org/10.3758/s13428-017-0971-x>.
- Fokkema, M., Edbrooke-Childs, J. & Wolpert, M. 2021 Generalized linear mixed-model (GLMM) trees: a flexible decision-tree method for multilevel and longitudinal data. *Psychotherapy Research* **31** (3), 329–341. <https://doi.org/10.1080/10503307.2020.1785037>.
- Fu, G., Jin, Y., Sun, S., Yuan, Z. & Butler, D. 2022 The role of deep learning in urban water management: a critical review. *Water Research* **223**, 118973. <https://doi.org/10.1016/j.watres.2022.118973>.
- Gelman, A. & Hill, J. 2007 *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY.
- Goldfarb, D. & Idnani, A. 1983 A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming* **27** (1), 1–33. <https://doi.org/10.1007/BF02591962>.
- Guimarães Santos, C. A. & Silva, G. B. L. d. 2014 Daily streamflow forecasting using a wavelet transform and artificial neural network hybrid models. *Hydrological Sciences Journal* **59** (2), 312–324. <https://doi.org/10.1080/02626667.2013.800944>.
- Guo, J., Sun, H. & Du, B. 2022 Multivariable time series forecasting for urban water demand based on temporal convolutional network combining random forest feature selection and discrete wavelet transform. *Water Resources Management* **36** (9), 3385–3400. <https://doi.org/10.1007/s11269-022-03207-z>.
- Hajjem, A., Bellavance, F. & Larocque, D. 2014 Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation* **84** (6), 1313–1328. <https://doi.org/10.1080/00949655.2012.741599>.
- Hastie, T. & Tibshirani, R. 1986 Generalized additive models. *Statistical Science* **1** (3), 297–310.
- Hastie, T., Tibshirani, R. & Friedman, J. 2009 *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd edn. Springer, New York.
- Hester, C. M. & Larson, K. L. 2016 Time-series analysis of water demands in three North Carolina cities. *Journal of Water Resources Planning and Management* **142** (8). Available from: <https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29WR.1943-5452.0000659> (accessed 7 June 2019).
- House-Peters, L. A. & Chang, H. 2011 Urban water demand modeling: review of concepts, methods, and organizing principles. *Water Resources Research* **47** (5). <https://doi.org/10.1029/2010WR009624>.
- House-Peters, L., Pratt, B. & Chang, H. 2010 Effects of urban spatial structure, sociodemographics, and climate on residential water consumption in Hillsboro, Oregon. *JAWRA Journal of the American Water Resources Association* **46** (3), 461–472. <https://doi.org/10.1111/j.1752-1688.2009.00415.x>.
- Hu, P., Tong, J., Wang, J., Yang, Y. & Oliveira, Turci L. de 2019 A hybrid model based on CNN and Bi-LSTM for urban water demand prediction. In *2019 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1088–1094. <https://doi.org/10.1109/CEC.2019.8790060>.
- Huang, H., Zhang, Z. & Song, F. 2021 An ensemble-learning-based method for short-term water demand forecasting. *Water Resources Management* **35** (6), 1757–1773. <https://doi.org/10.1007/s11269-021-02808-4>.
- Huang, H., Zhang, Z. & Song, F. 2022 A neural network approach for short-term water demand forecasting based on a sparse autoencoder. *Journal of Hydroinformatics* **25** (1), 70–84. <https://doi.org/10.2166/hydro.2022.089>.
- James, G., Witten, D., Hastie, T. & Tibshirani, R. 2021 *An Introduction to Statistical Learning*, 2nd edn. Springer. Available from: <https://www.statlearning.com> (accessed 24 February 2022).
- Kuncheva, L. I. 2014 *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons, Hoboken, NJ.
- Lamb, S. E., Haacker, E. M. K. & Smidt, S. J. 2021 Influence of irrigation drivers using boosted regression trees: Kansas high plains. *Water Resources Research* **57** (5). <https://doi.org/10.1029/2020WR028867>.

- Lee, S.-J., Chang, H. & Gober, P. 2015 Space and time dynamics of urban water demand in Portland, Oregon and Phoenix, Arizona. *Stochastic Environmental Research and Risk Assessment* **29** (4), 1135–1147. <https://doi.org/10.1007/s00477-014-1015-z>.
- Liaw, A. & Wiener, M. 2002 Classification and regression by randomForest. *R News* **2** (3), 18–22.
- Liemohn, M. W., Shane, A. D., Azari, A. R., Petersen, A. K., Swiger, B. M. & Mukhopadhyay, A. 2021 RMSE is not enough: guidelines to robust data-model comparisons for magnetospheric physics. *Journal of Atmospheric and Solar-Terrestrial Physics* **218**, 105624. <https://doi.org/10.1016/j.jastp.2021.105624>.
- Liu, G., Savic, D. & Fu, G. 2023 Short-term water demand forecasting using data-centric machine learning approaches. *Journal of Hydroinformatics* jh2023163. <https://doi.org/10.2166/hydro.2023.163>.
- McCarthy, M., Brogan, C., Shortridge, J., Burgholzer, R., Kleiner, J. & Scott, D. 2022 Estimating facility-level monthly water consumption of commercial, industrial, municipal, and thermoelectric users in Virginia. *JAWRA Journal of the American Water Resources Association* n/a (n/a). <https://doi.org/10.1111/1752-1688.13037>.
- Mini, C., Hogue, T. S. & Pincetl, S. 2014 Patterns and controlling factors of residential water use in Los Angeles, California. *Water Policy* **16** (6), 1054–1069. <https://doi.org/10.2166/wp.2014.029>.
- Mu, L., Zheng, F., Tao, R., Zhang, Q. & Kapelan, Z. 2020 Hourly and daily urban water demand predictions using a long short-term memory based model. *Journal of Water Resources Planning and Management* **146** (9), 05020017. [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0001276](https://doi.org/10.1061/(ASCE)WR.1943-5452.0001276).
- Pacific Institute 2013 *Water Rates: Water Demand Forecasting*. Available from: https://pacinst.org/wp-content/uploads/2013/01/water_rates_water_demand_forecasting.pdf.
- Perrone, D., Hornberger, G., van Vliet, O. & van der Velde, M. 2015 A review of the United States' past and projected water use. *JAWRA Journal of the American Water Resources Association* **51** (5), 1183–1191. <https://doi.org/10.1111/1752-1688.12301>.
- Polebitski, A. S. & Palmer, R. N. 2010 Seasonal residential water demand forecasting for census tracts. *Journal of Water Resources Planning and Management* **136** (1), 27–36. [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0000003](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000003).
- Sankarasubramanian, A., Sabo, J. L., Larson, K. L., Seo, S. B., Sinha, T., Bhowmik, R., Vidal, A. R., Kunkel, K., Mahinthakumar, G., Berglund, E. Z. & Kominoski, J. 2017 Synthesis of public water supply use in the United States: spatio-temporal patterns and socio-economic controls. *Earth's Future* **5** (7), 771–788. <https://doi.org/10.1002/2016EF000511>.
- Seibold, H., Hothorn, T. & Zeileis, A. 2019 Generalised linear model trees with global additive effects. *Advances in Data Analysis and Classification* **13** (3), 703–725. <https://doi.org/10.1007/s11634-018-0342-1>.
- Sharvelle, S., Dozier, A., Arabi, M. & Reichel, B. 2017 A geospatially-enabled web tool for urban water demand forecasting and assessment of alternative urban water management strategies. *Environmental Modelling & Software* **97**, 213–228. <https://doi.org/10.1016/j.envsoft.2017.08.009>.
- Shortridge, J. E., Guikema, S. D. & Zaitchik, B. F. 2016 Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences* **20** (7), 2611–2628. <https://doi.org/10.5194/hess-20-2611-2016>.
- Shortridge, J. & DiCarlo, M. F. 2020 Characterizing trends, variability, and statistical drivers of multisectoral water withdrawals for statewide planning. *Journal of Water Resources Planning and Management* **146** (3), 04020002. [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0001175](https://doi.org/10.1061/(ASCE)WR.1943-5452.0001175).
- Suero, F. J., Mayer, P. W. & Rosenberg, D. E. 2012 'Estimating and verifying United States households' potential to conserve water'. *Journal of Water Resources Planning and Management* **138** (3), 299–306. [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0000182](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000182).
- Tibshirani, R., Walther, G. & Hastie, T. 2001 Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63** (2), 411–423. <https://doi.org/10.1111/1467-9868.00293>.
- Toth, E., Bragalli, C. & Neri, M. 2018 Assessing the significance of tourism and climate on residential water demand: panel-data analysis and non-linear modelling of monthly water consumptions. *Environmental Modelling & Software* **103**, 52–61. <https://doi.org/10.1016/j.envsoft.2018.01.011>.
- Turlach, B. A., Weingessel, A. & Moler, C. 2019 quadprog: Functions to Solve Quadratic Programming Problems. Comprehensive R Archive Network. Available from: <https://cran.r-project.org/web/packages/quadprog/index.html>.
- Virginia Department of Environmental Quality 2022 *Virginia State Water Resources Plan: A Report of Virginia's Water Resources*. Richmond, VA, p. 627. Available from: <https://www.deq.virginia.gov/home/showpublisheddocument/13286/637781058061970000>.
- Vörösmarty, C. J., Green, P., Salisbury, J. & Lammers, R. B. 2000 Global water resources: vulnerability from climate change and population growth. *Science* **289** (5477), 284–288. <https://doi.org/10.1126/science.289.5477.284>.
- Willuweit, L. & O'Sullivan, J. J. 2013 A decision support tool for sustainable planning of urban water systems: presenting the dynamic urban water simulation model. *Water Research* **47** (20), 7206–7220. <https://doi.org/10.1016/j.watres.2013.09.060>.
- Wongso, E., Nateghi, R., Zaitchik, B., Quiring, S. & Kumar, R. 2020 A data-driven framework to characterize state-level water use in the United States. *Water Resources Research* **56** (9). <https://doi.org/10.1029/2019WR024894>.
- Wood, S. N. 2011 Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** (1), 3–36. <https://doi.org/10.1111/j.1467-9868.2010.00749.x>.
- Worland, S. C., Steinschneider, S. & Hornberger, G. M. 2018 Drivers of variability in public-supply water use across the contiguous United States. *Water Resources Research* **54** (3), 1868–1889. <https://doi.org/10.1002/2017WR021268>.

- Zanfei, A., Brentan, B. M., Menapace, A. & Righetti, M. 2022 A short-term water demand forecasting model using multivariate long short-term memory with meteorological data. *Journal of Hydroinformatics* **24** (5), 1053–1065. <https://doi.org/10.2166/hydro.2022.055>.
- Zhang, Z. & Balay, J. W. 2014 How much is too much?: challenges to water withdrawal and consumptive use management. *Journal of Water Resources Planning and Management* **140** (6), 01814001. [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0000446](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000446).

First received 27 April 2023; accepted in revised form 14 August 2023. Available online 31 August 2023