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

Julie Shortridge, Department of Biological Systems Engineering, Virginia Tech

Key points:

1. A novel method for withdrawal prediction using hierarchical machine learning is presented.
2. Hierarchical ensemble models reduce predictive errors for a majority of facilities analyzed.
3. Ensemble models are most beneficial in facilities with high variance and fewer observations of withdrawal.

Keywords: Water management (6334), human impacts (4323), anthropogenic effects (4802, 4902), machine learning (0555)

Abstract:

Accurate models of water withdrawal are crucial in anticipating the potential water use impacts of drought and climate change. Machine-learning methods are increasingly used in water withdrawal prediction due to their ability to model the complex, nonlinear relationship between water use and potential explanatory factors. However, most machine learning methods do not explicitly address the hierarchical nature of water use data, where multiple observations are typically available for multiple facilities, and these facilities can be grouped an organized in a variety of different ways. This work presents a novel approach for prediction of water withdrawals across multiple usage sectors using an ensemble of models fit at different hierarchical levels. A dataset of over 300,000 records of water withdrawal was used to fit models at the facility and sectoral grouping levels, as well as across facility clusters defined by temporal water use characteristics. Using repeated holdout cross validation, it demonstrates that ensemble predictions based on models learned from different data groupings improve withdrawal predictions for 63% of facilities relative to facility-level models. 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. Inspection of the ensemble weights indicated that cluster level weights were often higher than sector level weights, pointing towards the value of learning from the behavior of facilities with similar water use patterns, even if they are in a different sector.

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, particularly given the climatic sensitivity of outdoor and cooling water use. 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). Accurate models and projections of water demand are especially valuable in locations where water management institutions have relatively limited control on water use; this is the case in many Eastern states 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 (Perrone et al., 2015).

Recognizing this need, numerous studies have used statistical regression models to identify the environmental, socioeconomic, and institutional factors associated with greater volumes of water use. For instance, multiple studies have demonstrated the relationship between climatic conditions, land use, and household water use in specific municipalities (e.g., Balling et al., 2008; L. House-Peters et al., 2010; Lee et al., 2015; Mini et al., 2014). Several studies have leveraged broad-scale water use data to characterize drivers of broad geographic variability in per-capita municipal water use efficiency and trends (Chinnasamy et al., 2021; Sankarasubramanian et al., 2017; Worland et al., 2018) and irrigation withdrawals (Das et al., 2018).

Because the factors that influence water use tend to be highly complex and nonlinear, there has been increasing interest in the use of machine learning to model the relationship between water use and potential explanatory factors. For instance, Bolorinos et al. (2020) used random forests to model bi-monthly water demand for municipal customers and identify drought-induced changes in water consumption. Lamb et al. (2021) demonstrated how boosted regression trees could be used to model the relationship between groundwater pumping and numerical and categorical explanatory variables that can better represented governance differences. Recognizing that a large number of machine learning methods exist and may differ in terms of their predictive capabilities, Wongso et al. (2020) compared the ability of four different machine learning methods to predict state-level per-capita water use. When compared with linear regression approaches, ML models are often able to achieve lower predictive errors (Bolorinos et al., 2020; Wongso et al., 2020), pointing towards their potential value in water use modeling.

Across this body of research, one factor that is not always explicitly considered is the way in which the structure of water use data influences model predictions and inferences. Water use data is inherently hierarchical, with multiple potential options for grouping and categorizing observations. For instance, water use datasets often include observations through time for multiple customers or 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. Models of cross-sectional variability across different users and locations can assist in targeting conservation measures (Deoreo & Mayer, 2012; Suero et al., 2012), whereas models of temporal variability can lead to more accurate predictions of water use

under different policy and drought conditions (Hester & Larson, 2016). Recognizing this, longitudinal or mixed effects regression has become a standard statistical approach in modeling water use (Baerenklau et al., 2014; L. A. House-Peters & Chang, 2011; Polebitski & Palmer, 2010; J. Shortridge & DiCarlo, 2020). Longitudinal regression, and other forms of hierarchical regression more generally, allow model parameters to differ across groups in the data, while still constraining those parameter values based on population-level characteristics. 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).

Although hierarchical data structures have received less focus in the machine learning methodologies, recent research has begun to develop new approaches that account for grouped data structures. For example, the mixed effects random forest (MERF) approach models individual predictions through time as an additive function of a random forest model of population-level mean behavior processes and individual-level random effects (Hajjem et al., 2014). This approach was later extended to account for high-dimensional data that includes a large number of predictor variables relative to observations (Capitaine et al., 2021). Several studies have proposed methods that integrate regression and classification trees within a mixed modeling framework to address subgroups and hiearchies that exist in clinical trial data (Fokkema et al., 2018; Fokkema et al., 2021; Seibold et al., 2019). Other methods leverage ensemble learning, where predictions from multiple models are aggregated into a single prediction. For instance, Eygi Erdogan et al. (2021) create an ensemble of support vector machine models trained on different time intervals in a panel dataset. Ensemble learning, in which multiple models are independently fit to a dataset and combined into a single prediction,

has been found to generally reduce model variance which results in more accurate predictions on new data (James et al., 2021; Kuncheva, 2014).

The objective of this research was to develop and assess a novel approach for prediction of water withdrawals across multiple usage sectors using an ensemble of machine learning 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, and used multiple climatic and socioeconomic variables as predictor variables. A cluster analysis was also conducted to identify clusters of facilities with similar temporal patterns of water withdrawal, even if they were not in the same usage sector, with cluster-level models fit to observations within these groups. 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. In addition to assessing predictive accuracy, the relative weights applied to different grouping level models were used to evaluate the relative value of different grouping levels within the ensemble approach. 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

2.1.1 Withdrawal Data

This analysis used long-term records water withdrawal provided by the Virginia Department of Environmental Quality (VDEQ). All water users in Virginia who withdrawal more than 10,000 gallons per day (non-agricultural users) or 1 million gallons (MG) in any single month for crop irrigation 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). County location is available for all facilities within the dataset. Note that agriculture refers to livestock and agricultural processing operations, rather than crop irrigation. Additional details on withdrawal data are presented in Shortridge and 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.

	All Data		Retained for Analysis		
Sector	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
Total	*2,579*	*313,321*	*1,509*	*301,724*	*246,000*

Table 1: Summary of withdrawal data used in model development

Water withdrawal volumes across different users, even within a single usage sector, often vary across several orders of magnitude. Many water users also exhibit significant seasonal water user patterns as well. 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 two years of data for each facility.

2.1.2 Predictor Variables

A total of thirteen socio-economic and four climatic predictor variables were included as potential predictors of water withdrawal. Socio-economic variables are summarized in Table 2. These variables represented a variety of population, economic, and land-use characteristics that have been shown to have relationships with water withdrawals in previous research (e.g., Sankarasubramanian et al., 2017; Shortridge & DiCarlo, 2020; Worland et al., 2018). Socio-economic predictor variables were obtained from the U.S. Census (US Census Bureau, 2022), U.S. Bureau of Economic Analysis (US Bureau of Economic Analysis, 2022), USDA National Agricultural Statistics Service (US Department of Agriculture, 2017), the USGS FORecasting SCEnarios of Land-use Change (FORE-SCE) model (Sohl et al., 2007), and the U.S. Energy

Information Administration (US Energy Information Administration, 2022). Predictor variables that exhibited trends through time were linearly detrended prior to inclusion in the model (see Table 2).

Additionally, three climatic predictor variables were included to account for widespread evidence of the relationship between weather and water withdrawals (e.g., Brown et al., 2013; House-Peters & Chang, 2011; Lee et al., 2015). This included maximum daily temperature, total precipitation, and daily precipitation variability as quantified by the Gini coefficient (Marston & Ellis, 2019; Rajah et al., 2014). The Gini coefficient is a non-dimensional representation of the inequality in a distribution ranging from 0 to 1, with higher values indicating more inequality. Recent research has demonstrated an increasing trend in Gini coefficient rainfall values in the Eastern U.S., indicating that a larger portion of rainfall is occurring on a smaller number of days (Marston & Ellis, 2019; Rajah et al., 2014); this increase in rainfall variability is also correlated with greater water use (Shortridge & DiCarlo, 2020). Gridded daily values of maximum temperature and precipitation from 1990 – 2018 were obtained from the PRISM dataset (Daly et al., 2008) and spatially aggregated to county-level daily values. These county level values were then temporally aggregated to monthly estimate of average high temperature, total precipitation, and precipitation Gini coefficient.

Virginia exhibits several spatial patterns of variability in long-term climate conditions, most notably higher temperatures in the southern inland portion of the state, lower temperatures in the Appalachian and Blue Ridge Mountains, and a rain shadow along the western border of the state. To account for this cross-sectional variability as well as seasonal differences, all monthly climate data were transformed to monthly anomaly values prior to inclusion within the model (Shortridge et al., 2016):

$$AN_{c,t} = \frac{O_{c,t} - \bar{O}_{c,m}}{sd(O_{c,m})} \quad (2)$$

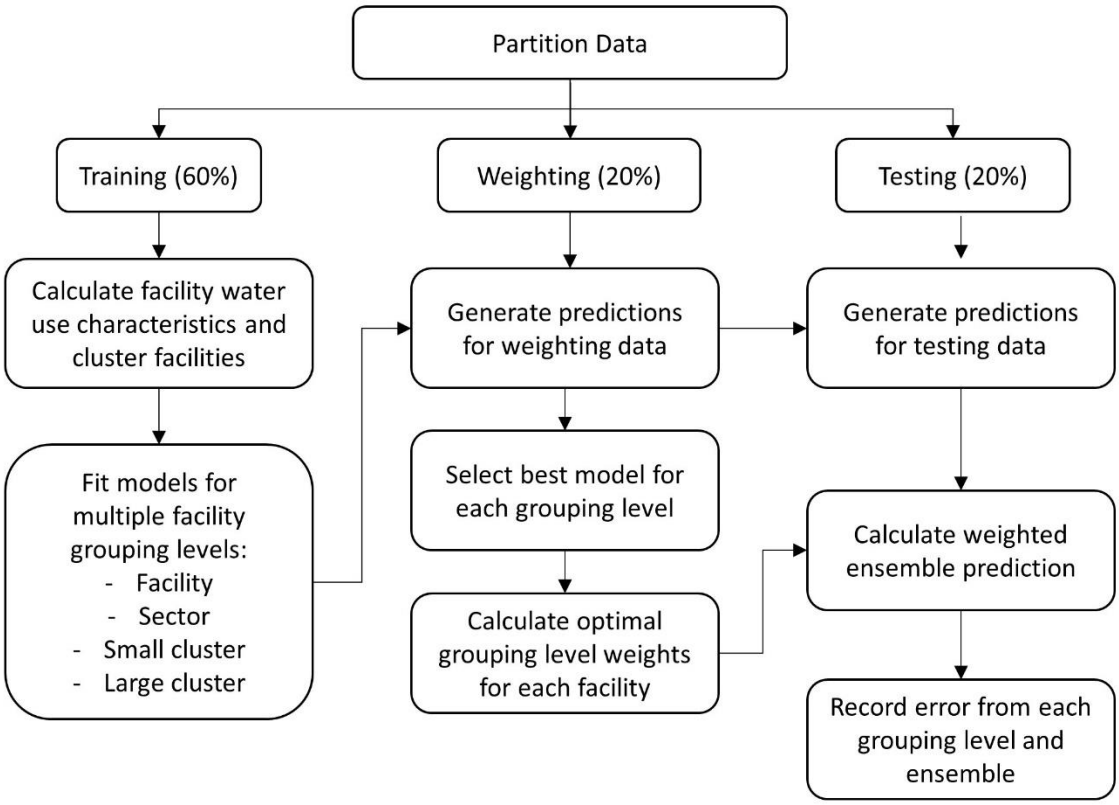
where $O_{c,t}$ is the observed value of a climate variable (e.g., total precipitation) in county c and time period t; $\bar{O}_{c,m}$ is the monthly normal (long-term mean from 1990-2018) observation of that variable in that county; and $sd(O_{c,m})$ is the standard deviation of all observations in that month. In this way, the anomaly value can represent how weather conditions compared to long-term averages for that specific month and location. For instance, an anomaly temperature value of 0.0 would indicate that the temperature in that month and county was exactly equal to the long-term mean temperature in that month and county, while a value of 1.0 would indicate that the temperature was one standard deviation higher than the long-term mean.

Explanatory Variable	Description and Rationale (temporal and geographic resolution in parentheses)	Source	Sector models
Population	County-level population (detrended); included to account for higher municipal water withdrawals as population grows (annual, county)	US Census	Mun
Per-capita personal income	Percent change (from previous year) in personal income per capita; Spatially-explicit representation of general economic conditions at the county level (annual, county)	US Bureau of Economic Analysis	All
Manufacturing GDP	Virginia GDP in current dollars from manufacturing (detrended); Spatially-general representation of manufacturing sector economic strength (annual, state)		Com, Ind, Min
Agricultural GDP	Virginia GDP in current dollars from agriculture (detrended); Spatially-general representation of agricultural sector economic strength (annual, state)		Ag, Aq, Com, Irr
USDA Price Ratio	Ratio of USDA Prices Received Index to Prices Paid Index (detrended); Spatially-general representation of economic conditions for agricultural producers (monthly, national)	USDA NASS	Ag, Aq, Irr
Agricultural Land Cover	Percentage of county land use classified as cultivated cropland, hay or pasture; Account for expansion and decline of agricultural land cover over study period (annual, county)	USGS FORE-SCE land cover model	Ag, Aq, Irr
Developed land cover density	Population per square kilometer of developed area; Included to account for differing development densities (annual, county)		Mun
Energy Prices	Virginia total energy prices in dollars per million BTU (detrended); Spatially-general representation of energy prices (annual, state)	US Energy Information Administration	All
Electricity Sales	Virginia retail electricity sales in millions of kWh (detrended); Spatially-general representation of energy consumption (annual, state)		Thm
Energy Production	Virginia primary energy production in trillion BTU (detrended); Spatially-general representation of energy production (annual, state)		Min, Thm

Table 2: Summary of socio-economic predictor variables

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. At each level, multiple model formulations (parametric linear models, semi-parametric non-linear models, and nonparametric machine learning models) 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.



211

212 *Figure 1: Diagram overview of modeling approach used during each iteration of cross*

213 *validation*

214

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.

Table 3: Summary of models and rationale for inclusion

2.2.1 Facility Grouping and Clustering

The water withdrawal data used in this study can be grouped at multiple different levels, some of which are hierarchical. 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 3). 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. However, it can result in a model that is 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 pattern of water use, particularly in the industrial and commercial sectors (Attaallah, 2018; McCarthy et al., 2022). Thus, the small and large cluster grouping levels were determined based 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 three-month mean withdrawal divided by the highest three-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.

Figures summarizing observed values for the five characteristics for all facilities are included as supplementary material. To determine the optimal number of clusters, facilities were divided into $k \in \{1, 15\}$ hierarchical clusters based on Euclidian distance using the FactoExtra package in R (Kassambara & Mundt, 2020). Gap statistic estimates for each value of k (included in the supplementary material) 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 while certain clusters largely corresponded to a single sector, the majority did not. This suggests that there are certain patterns of water use behavior that cannot be explained simply by sectoral classifications, echoing previous research that demonstrates significant variability across water use behavior across facilities that are generally grouped into a single sectoral classification (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 the supplementary material.

2.2.2 Regression and Machine Learning Models

For each of the grouping levels described above, multiple regression and 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 3, 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 x t matrix of m predictor variables across t months $\boldsymbol{X_c}$, plus an error term ε. Note that because predictor variables were available at the county rather than facility level, facility withdrawal is estimated as a function of predictor variables for its county location.

$$\boldsymbol{W\_AN_f} = f(\boldsymbol{X_c}) + \varepsilon \quad (3)$$

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-parameteric Gaussian generalized additive model (GAM), where smoothing functions are applied to the predictor variables to capture non-linear relationships between the 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 random forest (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 4 prior to estimating model error:

$$\boldsymbol{W\_P_f} = \boldsymbol{W\_AN_f}\, x\, sd\left(W\_O_{f,m}\right) + \overline{W\_O}_{f,m} \quad (4)$$

For each grouping level model, the GLM, GAM, and RF models were compared in terms of their mean absolute error across the weighting dataset, with the lowest error model retained. Following this process, each facility had four sets of withdrawal predictions 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:

$$\boldsymbol{W\_P\_Ens_f} = \boldsymbol{w_f} \cdot \boldsymbol{W\_P_{f\_all}} \quad (5)$$

Where $\boldsymbol{w_f}$ is a facility-specific vector of weights summing to 1.0, and $\boldsymbol{W\_P_{f\_all}}$ is an *n x 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 $\boldsymbol{W\_P\_Ens_f}$ is thus

a vector of *n* predictions obtained from a weighted average of individual grouping level model predictions. The values of the weights $\boldsymbol{w_f}$ for each facility were estimated using quadratic programming problem (Goldfarb & Idnani, 1983) implemented via the quadprog package in R (Turlach et al., 2019) of the form:

$$Minimize\ (\boldsymbol{W\_O_f} - \boldsymbol{w_f} \cdot \boldsymbol{W\_P_{f\_all}})^2 \quad (6)$$

Subject to

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

$$w_i \geq 0.0\ \forall\ i \in [1,4] \quad (7)$$

The four grouping level models, as well as the ensemble model based on the weights calculated in equations 6-7, were then used to generate withdrawal predictions for the testing dataset.

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. The training data were used to fit three models of different functional forms (GLM, GAM, and RF) for each grouping level described above. 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 determines the weights used in the ensemble model. Finally, the grouping level and ensemble models were used to predict withdrawals in the testing dataset. 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. Mean absolute error (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).

$$MAE_f = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{N}\sum_{n=1}^{N} abs(W\_O_{f,n} - W\_P_{f,n}) \quad (8)$$

$$RMAE_f = \frac{1}{H}\sum_{h=1}^{H}\frac{MAE_{f,h}}{\overline{W\_O}_{f,h}} \quad (9)$$

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 $\boldsymbol{C}$. This included the logarithm of mean faculty withdrawal, the number of nonzero observations, the coefficient of variation, autocorrelation, seasonality, and a use sector, as defined in Section 2.2.1 above.

$$EnsImp_f = RMAE_{f\_fac} - RMAE_{f\_ens} \quad (10)$$

$$EnsImp_f = \alpha + \boldsymbol{\beta}\, \boldsymbol{C_f} + \varepsilon \quad (11)$$

3. Results

3.1 Model Performance

A summary 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).

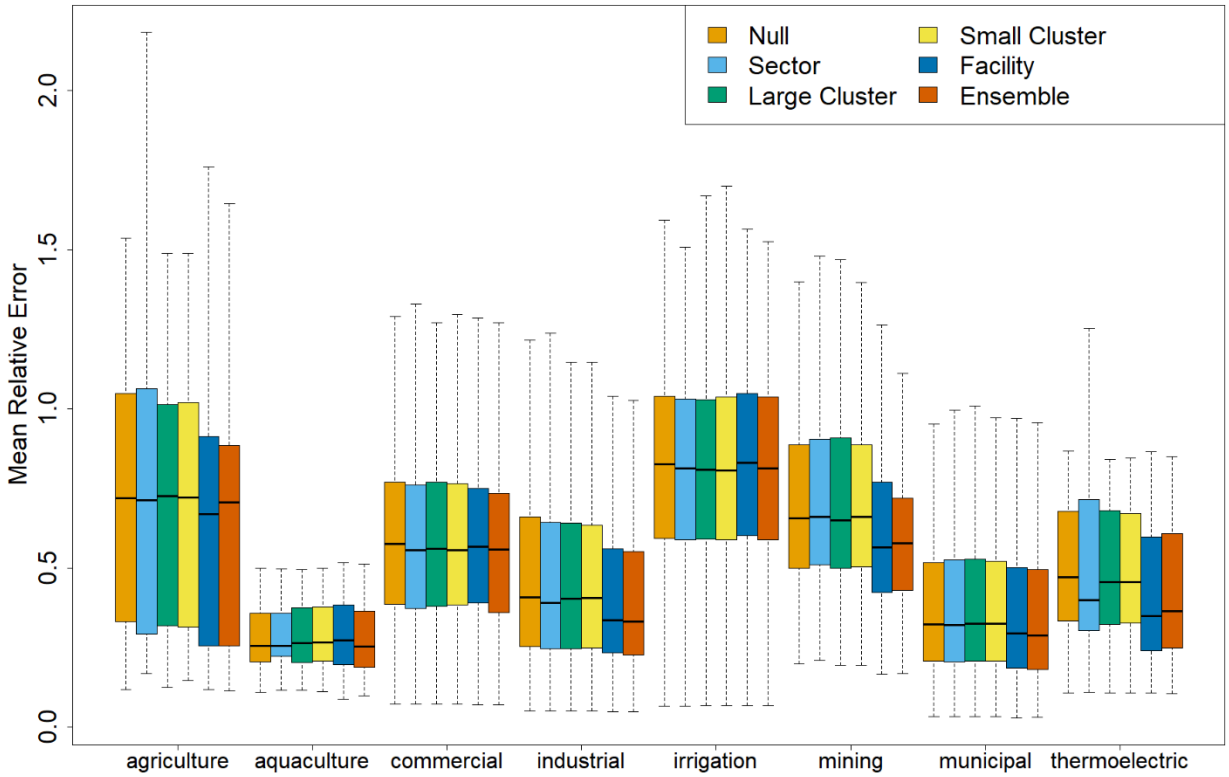


Figure 2: Relative percent errors across facilities in each water use sector. Boxplots show the distribution of RMAE for all facilities in each sector averaged across cross validation iterations.

Figure 3 presents a summary of the highest performing grouping level across facilities in each water use sector, as well as all facilities in the dataset. Across all facilities, the model form that most often resulted in lowest errors was the facility grouping, which had the best performance for 33.1% of the analyzed facilities, followed by the ensemble model (21.0%). On the other hand, the null model had the lowest errors in 15.9% of analyzed facilities. Facilities across different sectors exhibited different patterns in terms of optimal model selection. For instance, facility grouping models resulted in lowest errors in a large percentage of industrial (46.1%), mining (42.9%) and thermoelectric facilities (52.2%). However, facility models resulted in lowest errors in only 16.7%, 26.0%, and 27.8% of aquaculture, commercial, and

agricultural facilities, respectively. Sector-level models performed well for a relatively high percentage of irrigation facilities (lowest errors in 27.3% of facilities), but less well in the mining (7.1%), municipal (9.9%), and industrial (11.0%) sectors.

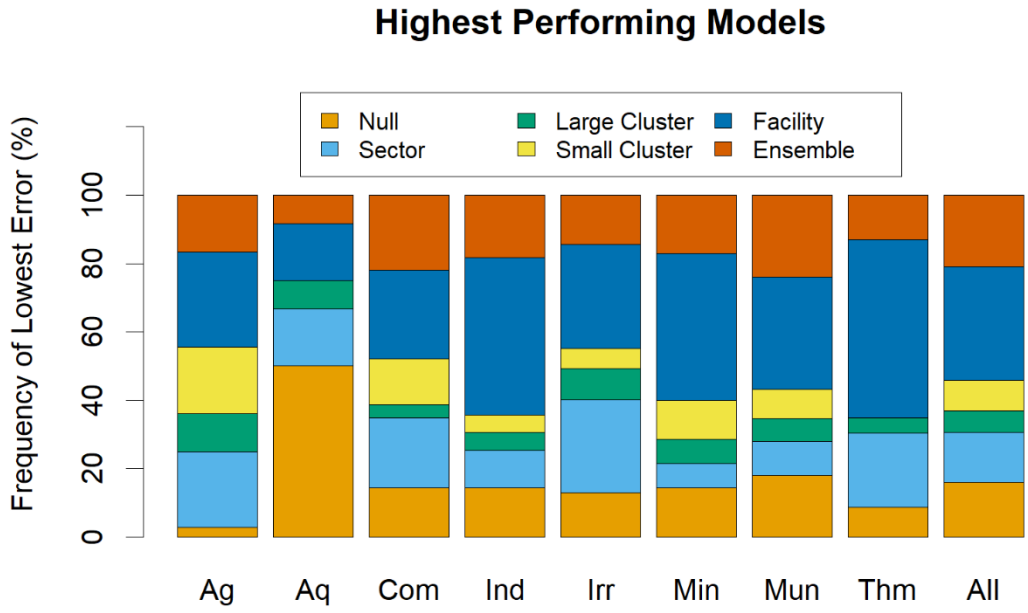


Figure 3: Percentage of facilities where each model type resulted in the highest performance in terms of RMAE.

Figure 4 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 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, with the ensemble approach reducing errors in 72.7% to 79.2% of facilities relative to the null, sector, and cluster models, but only 50% of facilities relative to facility-specific models. Ensemble improvements were most consistent in the municipal sectors, where errors were reduced by 61.5% to 68.3% of facilities, regardless of the grouping level model with which the ensemble is compared. The results in Figures 3 and 4 collectively demonstrate that while the ensemble process is unlikely to result in the optimum prediction for any single individual facility, it is likely to result in better predictions relative to single grouping level models when applied to many 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.

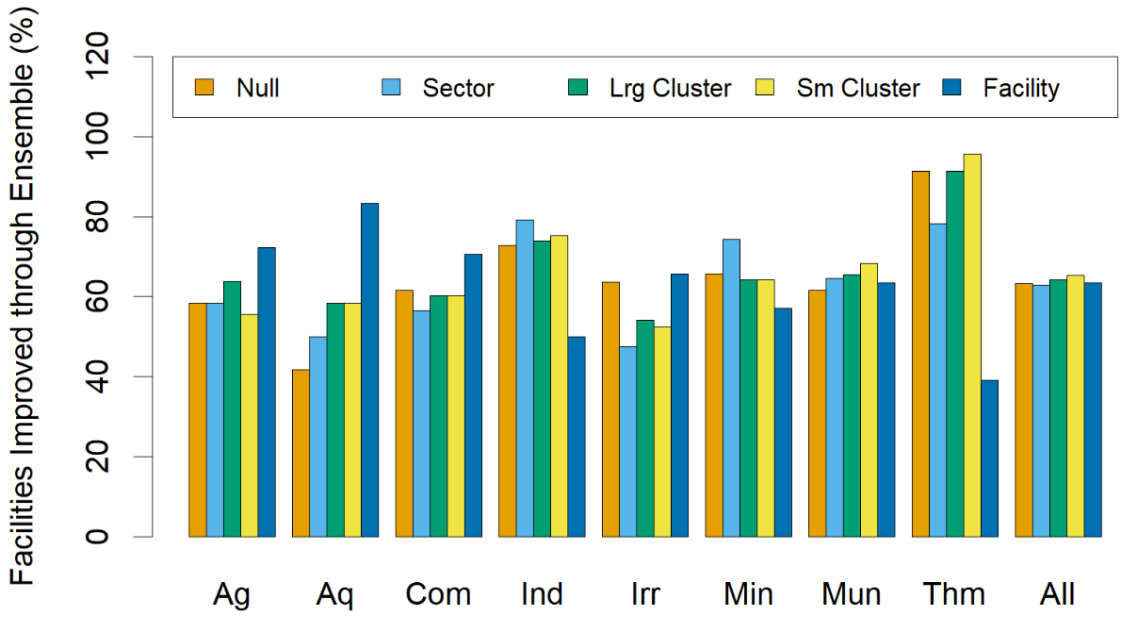


Figure 4: Percentage of facilities where ensemble model reduced errors relative to null and grouping level models

3.2 Ensemble Model Structure

To better understand the relative influence that different individual grouping level models play within the ensemble predictions, Figure 5 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, with the highest facility-level weights in the sectors where the facility models tended to perform best (industrial, mining, and thermoelectric). 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 demonstrates the value of learning from the behavior of facilities with similar water use patterns, even if they are in a different sector.

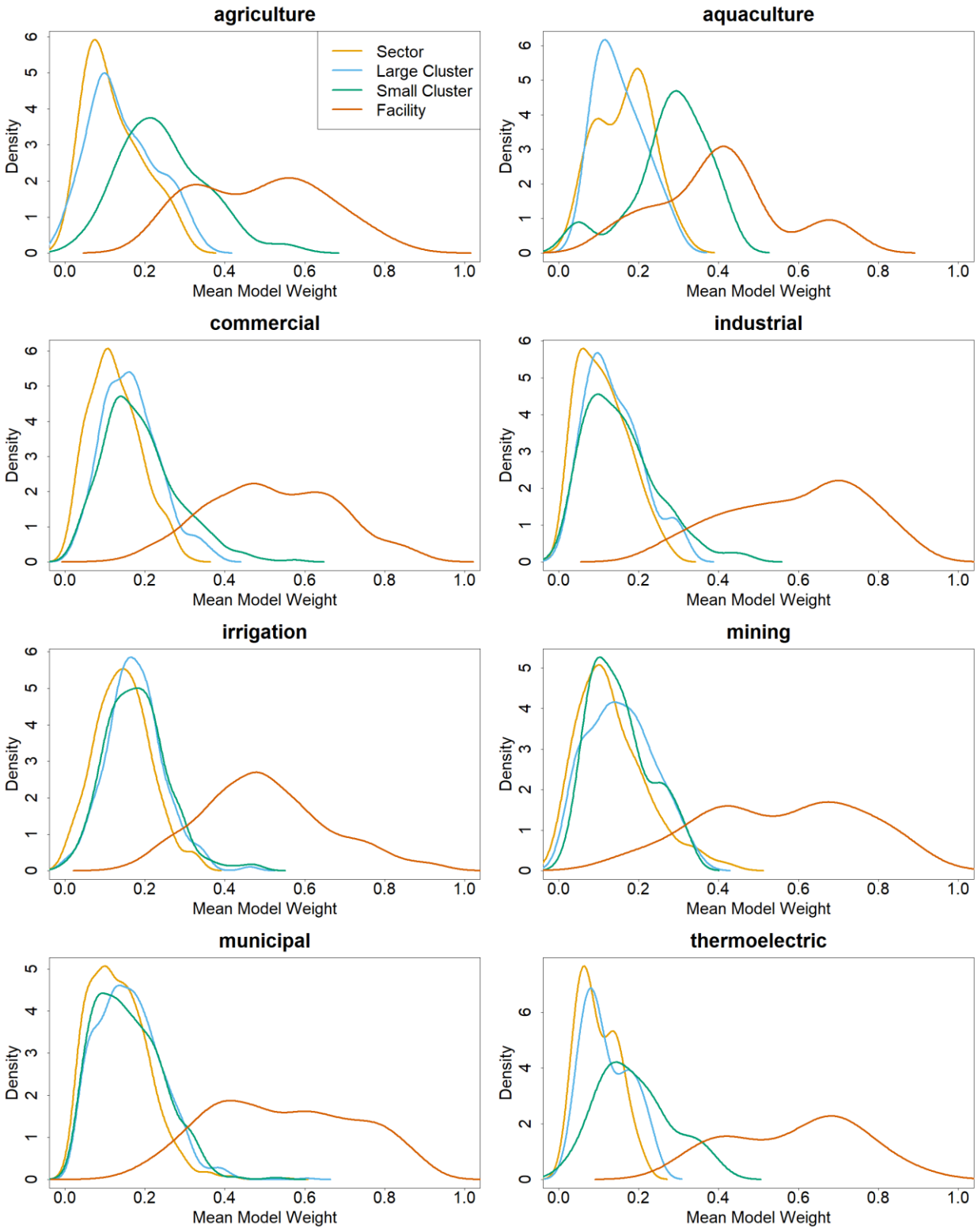


Figure 5: Density plot of mean ensemble model weights across facilities in each sector, averaged across cross validation iterations

	Estimate	Std. Error	p-value
Intercept	0.014	0.008	0.078
log(Water.Use.MGM)	-0.001	0.001	0.127
n.obs.nonzero	**-2.85E-05**	**1.13E-05**	**0.012**
Water.Use.COV	**0.016**	**0.001**	**< 0.001**
Water.Use.ACF.strength	**-0.022**	**0.004**	**< 0.001**
Water.Use.Seasonality	0.007	0.005	0.186
UseType (aquaculture)	0.003	0.013	0.839
UseType (commercial)	-0.001	0.007	0.832
UseType (industrial)	-0.005	0.008	0.466
UseType (irrigation)	**-0.016**	**0.007**	**0.026**
UseType (mining)	0.015	0.008	0.078
UseType (municipal)	0.004	0.007	0.603
UseType (thermoelectric)	-0.005	0.011	0.631

Table 4: Factors associated with improved ensemble performance relative to facility model performance. Bold rows indicate factors significantly associated with a difference in facility and ensemble performance ($p < 0.05$).

To better understand the facility water use 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 4. 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.

4. Discussion

Comparing the relative accuracy of different model grouping levels can provide some insights into the situations in which different model structures may provide optimal predictions. Across all facilities, the facility-grouping level model most frequently resulted in the lowest errors, but this still only occurred in 33.1% of facilities. Thus, there were numerous facilities

where a facility-specific model resulted in higher errors than other grouping levels. While the ensemble model only resulted in error minimization in 21% of facilities, it resulted in error reduction relative to grouping level models in over 60% of facilities. In this sense, its value is likely highest in situations where a general modeling approach is needed to simulate withdrawals across a heterogenous body of water users. The regression results in Table 4 indicate that this is particularly true in facilities with fewer observations, greater variance, and less autocorrelation. 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). High variance in water use observations could also potentially be indicative of data quality issues in self-reported water use data that have been observed elsewhere (Chini & Stillwell, 2017; McCarthy et al., 2022; Zhang & Balay, 2014). The ability to reduce the impact of these errors in predictive models is another potential benefit to the ensemble modeling approach.

This work also provides some insights into the relative abilities and limitations in applying sector-based models for facility-level predictions. Sector grouping models performed relatively well among agricultural and irrigation facilities, but less well in the industrial, mining, and municipal sectors. This is possibly due to the high variability in water withdrawal practices in those sectors. In particular, the industrial sector contains facilities that have a wide range of end uses for water, including cooling, incorporation into products, and landscape irrigation; this has been shown to lead to a high degree of variability across facilities in terms of water use and consumption rates (Attaallah, 2018; McCarthy et al., 2022). It should also be noted that a single municipal water withdrawing facility in our dataset could potentially serve multiple water supply utilities beyond the county in which the water is withdrawn. These water transfers present a

challenge for modeling and predicting withdrawal as water demand may be driven by conditions in counties other than the location of the withdrawal.

Several areas of additional research could be envisioned to build on the results presented here. For instance, we grouped facilities based on temporal water usage characteristics and sectoral classifications. However, water withdrawals could also depend on regulatory governance, with different withdrawal patterns expected in more strict regulatory environments (such as groundwater management zones). Water source could also be used as a grouping level, as surface water sources may experience more short-term fluctuations in water availability than groundwater. Exploration of alternative grouping strategies could be a valuable area of additional research. Similarly, this work used a global error metric (MAE) that aggregates predictive error across all observations equally to select the best performing model. However, as water supply management is typically more concerned with periods of stress and potential water shortage, other metrics that quantify model performance in terms of identifying periods of high withdrawal may be of value. For instance, event detection metrics quantify the degree to which the model captures specific conditions of interest, such as values above a predefined threshold (Liemohn et al., 2021). The concept of domain applicability can also be used to identify subdomains or conditions (such as high or low withdrawal periods) in which different models perform optimally (Sutton et al., 2020). The exploration of different performance measures for model selection would be a valuable area for further research.

5. Conclusions

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. Across all facilities, the model form that most often resulted in lowest errors was the facility grouping model (33.1%) followed by the ensemble model (21.0%). 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 towards the value of learning from 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 heterogenous, multi-sector groupings of water users.

6. Acknowledgements

I would like to gratefully acknowledge the Virginia Department of Environmental Quality for providing the data used in this project. All code and data used in this analysis are available at: https://osf.io/5pqvx/?view_only=a9b67a7867eb411585897076bf36a433 . This work was financially supported through institutional funding provided by Virginia Tech; this support is gratefully acknowledged.

7. References

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, Logan, Utah.

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

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., & Thiébaut, 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

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

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., et al. (2008). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*, *28*(15), 2031–2064. https://doi.org/10.1002/joc.1688

Das, P., Patskoski, J., & Sankarasubramanian, A. (2018). Modeling the Irrigation Withdrawals Over the Coterminous US Using a Hierarchical Modeling Approach. *Water Resources Research*, *54*(6), 3769–3787. https://doi.org/10.1029/2017WR021723

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, Sabine. (2011). *Cluster Analysis*. London, United Kingdom: John Wiley & Sons.

Eygi Erdogan, B., Özöğür-Akyüz, S., & Karadayı Ataş, 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

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (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, Marjolein, 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

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.

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

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* (Second). New York: Springer.

548 Hester, C. M., & Larson, K. L. (2016). Time-Series Analysis of Water Demands in Three North
549 Carolina Cities | Journal of Water Resources Planning and Management | Vol 142, No 8.
550 *Journal of Water Resources Planning and Management*, *142*(8). Retrieved from
551 https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29WR.1943-5452.0000659
552 House-Peters, L., Pratt, B., & Chang, H. (2010). Effects of Urban Spatial Structure,
553 Sociodemographics, and Climate on Residential Water Consumption in Hillsboro,
554 Oregon. *JAWRA Journal of the American Water Resources Association*, *46*(3), 461–472.
555 https://doi.org/10.1111/j.1752-1688.2009.00415.x
556 House-Peters, L. A., & Chang, H. (2011). Urban water demand modeling: Review of concepts,
557 methods, and organizing principles. *Water Resources Research*, *47*(5).
558 https://doi.org/10.1029/2010WR009624
559 James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical*
560 *Learning* (2nd ed.). Springer. Retrieved from https://www.statlearning.com
561 Kassambara, A., & Mundt, F. (2020). *Factoextra: Extract and Visualize the Results of*
562 *Multivariate Data Analyses. R Package Version 1.0.7.* Retrieved from https://CRAN.R-
563 project.org/package=factoextra
564 Kuncheva, L. I. (2014). *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley &
565 Sons.
566 Lamb, S. E., Haacker, E. M. K., & Smidt, S. J. (2021). Influence of Irrigation Drivers Using
567 Boosted Regression Trees: Kansas High Plains. *Water Resources Research*, *57*(5).
568 https://doi.org/10.1029/2020WR028867
569 Lee, S.-J., Chang, H., & Gober, P. (2015). Space and time dynamics of urban water demand in
570 Portland, Oregon and Phoenix, Arizona. *Stochastic Environmental Research and Risk*
571 *Assessment*, *29*(4), 1135–1147. https://doi.org/10.1007/s00477-014-1015-z
572 Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3),
573 18–22.
574 Liemohn, M. W., Shane, A. D., Azari, A. R., Petersen, A. K., Swiger, B. M., & Mukhopadhyay,
575 A. (2021). RMSE is not enough: Guidelines to robust data-model comparisons for
576 magnetospheric physics. *Journal of Atmospheric and Solar-Terrestrial Physics*, *218*,
577 105624. https://doi.org/10.1016/j.jastp.2021.105624

Marston, M. L., & Ellis, A. W. (2019). Change in the uniformity of the temporal distribution of precipitation across the MidAtlantic region of the United States, 1950-2017. *Climate Research*, *78*(1), 69–81. https://doi.org/10.3354/cr01561

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

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

Rajah, K., O'Leary, T., Turner, A., Petrakis, G., Leonard, M., & Westra, S. (2014). Changes to the temporal distribution of daily precipitation: Changing precipitation temporal patterns. *Geophysical Research Letters*, *41*(24), 8887–8894. https://doi.org/10.1002/2014GL062156

Sankarasubramanian, A., Sabo, J. L., Larson, K. L., Seo, S. B., Sinha, T., Bhowmik, R., et al. (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

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

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. *Hydrol. Earth Syst. Sci.*, *20*(7), 2611–2628. https://doi.org/10.5194/hess-20-2611-2016

Sohl, T. L., Sayler, K. L., Drummond, M. A., & Loveland, T. R. (2007). The FORE-SCE model: a practical approach for projecting land cover change using scenario-based modeling. *Journal of Land Use Science*, *2*(2), 103–126. https://doi.org/10.1080/17474230701218202

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

Sutton, C., Boley, M., Ghiringhelli, L. M., Rupp, M., Vreeken, J., & Scheffler, M. (2020). Identifying domains of applicability of machine learning models for materials science. *Nature Communications*, *11*(1), 4428. https://doi.org/10.1038/s41467-020-17112-9

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

Turlach, B. A., Weingessel, A., & Moler, C. (2019). quadprog: Functions to Solve Quadratic Programming Problems. Comprehensive R Archive Network. Retrieved from https://cran.r-project.org/web/packages/quadprog/index.html

United States Department of Agriculture. (2017). National Agricultural Statistics Service (NASS) Quick Stats Database. Retrieved March 1, 2018, from https://www.nass.usda.gov/Quick_Stats/

US Bureau of Economic Analysis. (2022). BEA Data. Retrieved June 24, 2022, from https://www.bea.gov/data

US Census Bureau. (2022). United States Census Data. Retrieved June 24, 2022, from https://data.census.gov/cedsci/

US Energy Information Administration. (2022). Data Tools, Apps, and Maps - U.S. Energy Information Administration. Retrieved June 24, 2022, from https://www.eia.gov/tools/index.php

Virginia Department of Environmental Quality. (2022). *Virginia State Water Resources Plan: A Report of Virginia's Water Resources* (p. 627). Richmond, VA. Retrieved 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

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

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