Forecasting Urban Water Demand in California: Rethinking Model Evaluation

Steven Buck
University of California, Berkeley
Department of Agriculture and Resource Economics
stevenbuck@berkeley.edu

Hilary Soldati
University of California, Berkeley
Department of Agriculture and Resource Economics
soldati@berkeley.edu

David L. Sunding
University of California, Berkeley
Department of Agriculture and Resource Economics
sunding@berkeley.edu

Selected Paper prepared for presentation at the 2015 Agricultural & Applied Economics Association and Western Agricultural Economics Association Annual Meeting, San Francisco, CA, July 26-28

Abstract

Urban water managers rely heavily on forecasts of water consumption to determine management decisions and investment choices. Typical forecasts rely on simple models whose criteria for selection has little to do with their performance in predicting out-of-sample consumption levels. We demonstrate this issue by comparing forecast models selected on the basis of their ability to perform well in-sample versus out-of-sample. Our results highlight the benefits of developing out-of-sample evaluation criteria to ascertain model performance. Using annual data on single-family residential water consumption in Southern California we illustrate how prediction ability varies according to model evaluation method. Using a training dataset, this analysis finds that models ranking highly on in-sample performance significantly over-estimated consumption $(10\% - 25\%)$ five years out from the end of the training dataset relative to observed demands five years out from the end of the training dataset. Whereas, the top models selected using our out-of-sample criteria, came within 1% of the actual total consumption. Notably, projections of future demand for the in-sample models indicate increasing aggregate water consumption over a 25-year period, which contrasts against the downward trend predicted by the out-of-sample models.

1 Introduction

Accurate forecasts of urban water demand provide valuable information to water resource managers, whether in determining efficient pricing and allocation strategies or in evaluating the benefits of infrastructure improvements and expansion. In regions that face uncertainty in surface water supply reliability, accurate projections of anticipated future demand facilitates cost and benefit analysis of management tools and budgetary decisions. In the context of urban water demand, forecasting methods require the researcher to make assumptions about specific price measures (marginal versus average), functional form (log versus linear), and relevant demand determinants (income, lot size); these have been debated in the academic literature, though there is no consistent set of assumptions (Arbués, Garcıa-Valiñas, and Martınez-Espiñeira, 2003). Moreover, water managers may rely on in-sample assessments of fitness in selection of benchmark models, such as R^2, to be used for predicting long run expectations of residential water demand. However, such in-sample selection criteria do not reflect the underlying objective that motivates formation of demand forecasts, which requires a model that accurately predicts future demand. In this paper, we present an alternative urban water forecasting method, similar to that developed by Auffhammer and Carson (2008) and Auffhammer and Steinhauser (2012), which minimizes sensitivity to a priori structural arguments and employs out-of-sample selection standards, to predict urban water demand. We limit our analysis to the single family residential sector for illustrative purposes. The analysis can be easily extended to model consumption forecasts for other sectors of urban water demand (e.g. multi-family residential, commercial and industrial, and institutional demands). Our results suggest that the standard approach, which relies on in-sample model evaluation, yields projections that are significantly flawed, relative to models that are chosen based upon our out-of-sample assessment methods.

For the state of California, known for its vast network of developed water supplies, construction of sound estimates of future single family residential (SFR) water demand aids decision-making for water managers, policy makers, and other relevant stakeholders. Moreover, anticipated hydrolog- ical change, which suggests more frequent and more severe drought periods for the state, magnifies the importance of valid demand forecasts (Diffenbaugh, Swain, and Touma, 2015; Swain et al., 2014). The drought conditions of 2012-2015 included the state's year of lowest recorded precipitation and have resulted in unprecedented, mandatory reductions for residential water utilities across the state (Executive Office, 2015). In addition to high variability in precipitation, the dynamics of the state's surface water resources is often characterized by a pattern of spatial and temporal mismatch between supply and demand. Abundant winter precipitation, in the sparsely populated north, must be stored and redirected to the more densely populated south, which is supplemented by other regional imports. Environmental regulation of surface water flows brings additional supply uncertainty to the utilities that are challenged with meeting residential demand in their service areas. Demand that is unmet by these variable supplies impacts urban utilities and communities through increased pressure on groundwater supplies, purchases of high-cost water transfers, mandated conservation or rationing, and reliance on state and federal drought relief. Hence, more accurate urban demand forecasting reduces one element of uncertainty and assists water managers in their optimization problem, possibly evading costly short term solutions.

While water managers recognize the role water demand forecasts have on their management decisions, they continue to rely on simple models whose criteria for selection has little to do with their performance in predicting out-of-sample consumption levels. This is evident in planning documents such Urban Water Management Plans, which all 400+ urban water utilities in California are required to submit every five years. Using the SFR sector as an example, we highlight the benefits of developing out-of-sample evaluation criteria to ascertain model performance and to select models for developing forecasts of future demands.

The paper proceeds by to the following structure. Section 2 explores the relevant existing literature, including discussion of the theoretical arguments that motivate household water demand. In Section 3, features of the underlying data are summarized and reviewed. We explain our methodology in Section 4, which is followed by construction of our out-of-sample evaluation criteria in Section 5. Results are presented in Section 6, which includes discussion both of out-of-sample model prediction ability and of projections of future demand using our models selected according to in-sample and out-of-sample evaluation methods. Section 7 concludes and includes suggestion for further work.

2 Background

There exist a variety of methods and models for forecasting urban water demand in the academic and professional literature, including the use of artificial neural networks, time series analysis, simulation, and multivariate regression. Artificial neural networks and time series analysis are often favored in short-term demand forecasts, largely because they outperform other methods in the near term in terms of forecasting levels of demand (Herrera et al., 2010; Zhou et al., 2000). Their use for near-term demand modeling is not surprising since they rely heavily on the recent observed past to predict the near future. The main drawback of these models is that they often perform well for very short prediction periods (e.g. a few days or a month). In contrast, simulation based models such as Demand-Side Management Least-Cost Planning Decision Support System[1] (DSS) assume an underlying economic behavior to make predictions about future consumption. The drawback of these models is that the assumed economic behavior is usually hypothetical instead of being based on observed responses to demand factors.

Auffhammer and Steinhauser (2012) demonstrate the benefits of developing CO_2 emissions forecasts based on out-of-sample performance as opposed to in-sample performance

[1] Created by Maddaus Water Management, Alamo, California.

measures. In the spirit of their work, we compare forecasts of single-family residential water demand derived from models with top performing in-sample scores versus models with top performing out-of-sample scores.

3 Data

This analysis makes use of annual, retailer-level panel data on average monthly water consumption and relevant determinants, between 1994 and 2009, for SFR consumers in Southern California. In particular, this data represents a subset of the Metropolitan Water District's (MWD) consumer base, which is comprised of 26 member agencies. These agencies may offer services through a secondary retailer or directly to households, in which case the retailer is defined as the agency itself. In addition to average monthly consumption, given in hundreds of cubic feet (CCFs), we have collected data on two categories of price measures - marginal price and average price. Our marginal price variable is set by the median tier rate for the relevant retailer, while total average price represents the average total water bill over average total consumption. Retailer data also includes yearly account totals.

In addition, our dataset includes several determinants of residential water demand, as given by the relevant literature. Data on household characteristics are taken from the U.S. Census and DataQuick and averaged to the retailer-level. These household attributes are: average lot size, average household size, and average income. We also include the following environmental drivers of residential demand: average temperature maximum, average summertime temperature maximum, and precipitation. Table 1 provides sum- mary statistics of our underlying data. We observe substantial variation across our variables.

It is important to note some challenging features of the data collection process. First, while all MWD agencies are represented in our analysis, data was not collected for the full subset of retailers within each agency. Constraints on collection and availability of data prevent inclusion of all retailers. Of the approximately 150 retailers who distribute MWD water, 98% of which is for Southern Californians, 113 are represented in this analysis.[2] However, the retailers for whom data collection was not feasible represent a relatively small portion of aggregate MWD consumption, both in total consumption and in number of accounts. For example, in 2005, the retailers represented in our dataset accounts for approximately 80% of the total MWD accounts and 90% of demand in that year, while this total constitutes 55% of the total retailers Thus, we have that our percent of retailers explains a larger proportion of total accounts. As is indicated in Table 1, the evident variation in the number of accounts per retailer reflects the fact that retailers vary substantially in scale of service area and total water consumption. This difference will be relevant for how models are scored in criteria that make use of forecasted values. We address this challenge in greater detail below in the section on score criteria and in our review of results.

A second challenge is that our panel is unbalanced in retailers across years. Heterogeneity in administrative organization and sophistication across retailers results in differences in the level and consistency of accessible data by retailer. Similarly, improvements in administrative organization, which correlates positively with time, imply greater availability of data in the latter portion of our panel dataset, relative to earlier years. Thus, in some years, data is unavailable for some retailers, with a higher probability of missing data in earlier years. In Figure 1, we plot changes in the number of retailers and in the total number of accounts through time that is available for our analysis. We see that, by 1999, the level of both retailers and number of accounts represented in our panel has achieved a degree of consistency, with an average of 88 retailers, servicing an average of 2.3 million accounts, the years between 1999-2009, inclusive.

To further investigate the nature of data consistency, we evaluate the number of retailers

[2] In total, MWD holds contracts with approximately 190 retailers. However, nearly 40 of these retailers are for unusually small service areas, sometimes serving just a couple of accounts. There are around 153 retailers serving more than 3,000 accounts, which we consider a more accurate figure when estimating the number of relevant MWD retailers.

for which we have different amounts of years available. Table 2 summarizes this exploration. While only 13% of our retailers have data available in all years of our panel, nearly 62% of retailers have more than 10 years of data available. Again, this concern will be most relevant in assessing models along criteria that use predicted consumption, which we discuss below.

4 Model Universe

In this analysis, we take a unique approach in our model of residential water demand. Rather than commit to a particular model of demand that represents a single theory of household consumption, we develop a flexible, computationally-driven process that minimizes the number of required assumptions. Following both the academic literature, as well as benchmark models used to predict future demand, we include both the full suite of demand determinants, that are understood to be drivers of residential water consumption, and several functional specifications in framing our model universe. However, we also attempt to strike a balance between full permutation of all possible models and establishing some broad restrictions that conform to theoretical standards. This process results in a sizable model universe, which is subject to both in-sample and out-of-sample evaluation methods.

This first step in establishing our model universe is to consider the possible covariate combinations that contribute to household demand. To begin, we consider a basic underlying regression specification:

$$q_{rat} = \beta price_{rat} + f(hhld_{rat}) + g(weather_{rat}) + h(q_{rat-j}) + k(time_{rat}) + \alpha_a + \epsilon_{rat} \qquad (1)$$

where q_{rat} is average monthly household consumption for retailer r, in agency a, for year t. We allow our price variable to take on the value of either the median tier rate (mtr_{rat}) or total average price (tac_{rat}), restricting our universe to models that do not include both price measures. Different possible household characteristics are represented by the vector $hhld_{rat}$, which includes a component for average household size (ahs_{rat}), median lot size, (mls_{rat}), and median household income (mhi_{rat}). When included, these covariates are assumed to enter the model linearly over all possible permutations of the three variables. Similarly, $weather_{rat}$ is a vector of retailer-level average weather characteristics, which are: precipitation ($prec_{rat}$), average maximum temperature ($tmax_{rat}$), and average summertime maximum temperature ($stmax_{rat}$). These variables also are given a linear specification, when included in a given model. However, here, we prohibit both $tmax_{rat}$ and $stmax_{rat}$ from ap-pearing together in any given specification. Also consistent with the existing literature, we allow for the possibility of lagged consumption, q_{rat-j} with $j \in (1, 2)$; we require inclusion of q_{rat-1} in models that use q_{rat-2}. Following Auffhammer and Steinhauser (2012), we consider a time trend, rather than time period fixed effects. We allow flexibility in our paramaterization of $time_{rat}$, up to a third degree polynomial. Lastly, we incorporate the option of agency fixed effects, α_a, as a possible determinant of demand.

Following the restrictions outlined above, we develop a model universe of $3,432$ total models. This set is formed by fully permuting over the covariate combinations described above. We further expand the universe of models by representing variables in both levels and under a natural log transformation. Moreover, each model specification is regressed employing three estimates: Ordinary Least Squares (OLS), weighted least squares (WLS) using an observation's proportion of total accounts as weights, and a robust regression estimator (RRE). Thus, our model evaluation process includes a total of $20,592$ empirical specifications.

5 Score Criteria

The objective of this research is to consider how accuracy of forecasted SFR water demand may be sensitive to biases toward particular model selection criteria. Such biases may erroneously emphasize models or establish benchmarks which do not prioritize actual forecast objectives in model selection. In particular, in-sample measures of model performance, such as R^2 and AIC, will prefer models that best fit *existing data*, rather than providing meaningful forecasts of out-of-sample consumption. Notably, these models will mechanically tend to prefer fixed effects methods. Following the approach of Auffhammer and Carson (2008) and Auffhammer and Steinhauser (2012), we develop three out-of-sample score criteria that correspond with three distinct forecasting objectives to evaluate model performance.

Before defining these out-of-sample criteria, we elaborate on the prediction process. To conduct our out-of-sample evaluation, we segment the underlying data into a *training set* and a *prediction set*. The training set is truncated at year t, while the prediction set includes observations for only year $p = t + \gamma$, for a γ-year prediction. For example, to evaluate a model's ability to accurately generate a 5-year prediction of consumption, we set $\gamma = 5$. In this instance, to maximize the availability of training data, we let $p = 2009$, which gives $t = 2004$. For each regression model in $m = 1, ..., 20, 592$, we use the estimated regression coefficients, which are derived using the training set, to generate predicted values in the $t + \gamma$ year.

To evaluate these predicted values, we employ three different measures of mean square forecast error, corresponding to three levels of observation: retailer, agency, and aggregate. These three levels of evaluation reflect different potential water management objectives. For instance, when forecasts are intended to shape retailer-level decisions, then we argue that evaluation criteria that minimizes the square forecast error at the retailer-level is most appropriate. On the other hand, when policy and budgetary decisions are over a larger system and of a broader scope, prioritizing model performance in aggregate forecast error would be preferred. A parallel argument may be applied to agency-level forecast objectives. As such, we propose three measures of out-of-sample performance: Retailer-Level Mean Square Forecast Error (RL-MSFE); Agency-Level Mean Square Forecast Error (AL-MSFE); and Aggregate

Forecast Error (AFE).

To obtain the RL-MSFE, for retailer r, we first generate predicted average consumption values, $\hat{q}_{rapm}$, using the regression coefficients estimated under model specification m, where p is our prediction year. We then create average annual consumption, using the corresponding number of accounts for retailer r, in year p, denoted as $\hat{Q}_{rapm}$. This prediction for retailer r's total consumption in year p is used to calculate the forecast error for each retailer, $(\hat{Q}_{rapm} - Q_{rapm})$. The square of these retailer-level forecast errors are summed and averaged, resulting in an RL-MSFE for model m. Thus, we have that:

$$RL - MSFE_m = \frac{\sum_{r=1}^{R} (\hat{Q}_{rapm} - Q_{rapm})^2}{R} \tag{2}$$

Formation of the AL-MSFE follows a similar logic, where forecast error is calculated at the agency-level. Thus, $\hat{Q}_{apm}$ is the predicted annual consumption for agency a in prediction year p, under model m. As above, we have:

$$AL - MSFE_m = \frac{\sum_{a=1}^{A} (\hat{Q}_{apm} - Q_{apm})^2}{A} \tag{3}$$

This method of evaluating models would be preferred in a setting where management decisions are being evaluated at the level of the member agency.

Finally, we may instead consider the aggregate forecast error, which keeps in mind a different policy and loss function. Rather than find the model that minimizes our loss function at the agency-level, one might prefer to measure prediction performance in the aggregate, over an entire region or service area. Hence, we develop a third out-of-sample method for evaluating model prediction performance, Aggregate Forecast Error (AFE):

$$AFE_m = \sum_{a=1}^{A} \hat{Q}_{apm} - \sum_{a=1}^{A} Q_{apm} \tag{4}$$

In plain language, equation 4 allows us to assess how far off in absolute terms our aggregate forecast in the year 2009 was relative to observed aggregate demand in 2009.

Broadly, there are two general approaches that may be used when considering a model's ability to produce accurate predictions. The first approach asks how a model performs in a γ-year prediction by creating a set of predicted values that are γ-years beyond the training set. Under such a setting, the researcher would repeat the γ-year prediction process, described above, π times by systematically stepping back the training set and the prediction set to generate π sets of predicted values. By repeating this process π times, the research avoids anomalous features of a particular training and prediction set and, instead, finds the model that produces the best γ-year prediction on average. This method, however, is constrained by the time horizon of the baseline data. While increasing π improves selection of the model that performs best, on average, for a γ-year prediction, the researcher is limited in increasing π by the size of the panel. For small datasets, a preferred choice of π may not be feasible as the training set becomes smaller and loses power to generate meaningful regression estimates.

A second approach would be to evaluate a model's prediction performance over different values of γ, which assesses a model's prediction ability in general, rather than for a specific value of γ. In this setting, the researcher would use the out-of-sample score criteria to rank models for each γ-year prediction, where $\gamma = 1....\Gamma$. The researcher then ranks the models according to performance on average, across the Γ prediction scores. Depending on the forecast objective or policy question to be answered, these scores may be given equal weights, for all Γ scores, or may be weighted to reflect analytic goals. For instance, a researcher may choose to place more weight on model rankings for larger prediction intervals. This framework finds the model that best predicts, on average, for any $\gamma \in \Gamma$ prediction length. This structure may be preferred when there is uncertainty over the prediction period in which policy makers or administrators are interested.

It is under this rubric of out-of-sample score criteria that limitations in our data take on significance. For this analysis, the short panel length restricts our choice of π and the unbalancedness in our panel affects forecast error. When our prediction years do not have data available for a consistent set of retailers, this may impact model ranking. To ameliorate this complication, we remove retailers for which we do not have data for all relevant prediction years. However, we allow our training set to remain unbalanced in order to maximize the available data for estimating coefficients of our explanatory variables.

6 Results

For this analysis, we rank model performance across four in-sample score criteria - R^2, Adjusted R^2, AIC, and BIC - and our three out-of-sample score criteria. Given our data limitations in this initial stage of research, we allow our models to make a 5-year prediction of average household consumption. Following the prediction process outlined above, we set $t = 2004$ for our training set, with $p = 2009$ as our prediction set. For each evaluation method, all models are ranked according to these score criteria. Additionally, we consider subcategories of models to understand how those different classes of models perform and may drive overall rankings within a score criteria. Predictions that are evaluated using logged data are treated with a Goldberger correction on the predicted values, such that:

$$\bar{q}_{ra(t+\gamma)m} = \bar{\sigma}_{rapm} e^{\hat{q}_{rapm}} \tag{5}$$

where $\bar{\sigma}_{rapm} = e^{\frac{\sigma_{\varepsilon}^2}{2}}$ and σ_{ε}^2 is the variance of the regression error term.

6.1 Model evaluation of 2009 prediction ability

Table 3 provides summary statistics on the top 1% of models for each score category, including the model subcategories of: levels only, log-log models only, exclusion of models with lagged consumption, exclusion of models with agency fixed effects. Thus, if we consider a particular score criteria, the first column reports the average value of that evaluation method for all the models that are in the top 1% of models, according to that score category. For example, the top 1% of all models using R^2 score criteria have an average R^2 value of 0.890. If we move across the R^2 row, we observe that, within the top 1% of models according to R^2 ranking, the log-log models perform the best. In the last column, we see that the average R^2 drops to 0.614 for models that do not include lagged consumption of the top 1% of models, as ranked by R^2. Similarly, logged models perform best by AIC criteria, while models that do not allow lagged consumption are significantly worse by this measure.

Our out-of-sample scores follow a similar pattern, with log-log specifications performing better than those given in levels and a significant increases in prediction error once lagged consumption is omitted as a covariate. This result may be driven by the fact that, over our sample period, aggregate consumption is non-parametric with respect to time. We see, in Figure 2, that aggregate annual consumption is first increasing in our sample and then decreases for the last several years. We observe the same pattern when we restrict our data to only include retailers that have at least twelve years of data, suggesting that this consumption trend is not being driven by the availability of data. Therefore, it is reasonable that models that are fitted with lagged consumption may over predict 2009 meter-use.

In Table 4, we explore how these models perform relative to AFE. For each model in the top 1% of models by each score criteria, we also evaluate that model's AFE and compute the average AFE for that set of models. This allows us to compare average model performance across a common forecasting score category. Thus, for each score category, we report the average AFE, in millions of CCFs, for the top 1% of models, according to that score criteria. For further insight, we also report the percent of actual 2009 aggregate demand, which was over 400 millions CCFs, that this average AFE represents, given in brackets. These results indicate that models which ranked highly according to both R^2 and Adjusted R^2 over predict 2009 consumption by a substantial amount. While all models in the top 1% by R^2 ranking over predict aggregate 2009 consumption by about 25%, this over prediction is driven largely by models that use a levels-levels specification. This relationship is consistent across all score categories. Across the top models in our two other in-sample categories, AIC and BIC, there is an improvement in prediction accuracy. However, these models are still over predicting, on average, by 15% and 10%, respectively.

As expected, we observe a substantial improvement in prediction ability for models that are chosen according to out-of-sample criteria. Models that are ranked according to RL-MSFE, our smallest spatial unit of observation, are slightly more precise than those models that rank highest by AL-MSFE criteria. Notably, the RL-MSFE models tend to slightly under predict consumption, while those that aggregate to the agency-level over predict consumption, on average. Unsurprisingly, the models that are chosen according the AFE criteria dramatically improve 2009 consumption predictions. The top 1% of models, by this evaluation category, tend to slightly under predict demand, which is driven by the log-log subclass of models.

Providing visual support for the results in Table 4, we plot predicted aggregate consumption for the top 1% of models, by each score criteria, from the training year (2004) through the desired prediction year (2009) in Figure 3. These plots include actual consumption (in blue), taken from our panel data, and the model that was ranked "best" by the relevant score criteria (in red). The fan of prediction paths (in grey) represent each model that is included in the top 1% of models by that score category. For instance, in Figure 3a, we that the majority of the models that rank highly by R^2 evaluation over predict 2009 consumption, confirming the results presented in Table 4. Again, we graphically confirm the improved performance for AIC models overall, and note that the "best" AIC model substantially improves the 2009 prediction, relative to R^2.[3]

We present the same graphical interpretation for all three of our out-of-sample score categories in Figure 4. These plots provide a stark visual argument for the claim that

[3] Please refer to the appendix for similar plots of both the *Adjusted* R^2 and BIC score criteria

using in-sample model selection methods may not yield desired outcomes for reliable demand forecasts, when compared to models that are selected based on out-of-sample performance.

6.2 Projections of future consumption

To demonstrate an application of these model selection methods, we generate forecasts of residential demand up through 2035. To produce these results, we require projected values for our covariates. These were made available through planning documents and datasets from the Southern California Association of Governments (SCAG) and the San Diego Association of Governments (SANDAG). These projections occur in 5-year intervals, beginning with 2010 and ending in 2035. For each model, we apply the coefficient estimates that result from regressing that model on the entire panel set to each year of covariate projections.

Summary statistics of projected aggregate demand are given in Table 5, after omitting models that allow for lags or higher order time trends, which are ordered according to performance within each evaluation method. For each score criteria, we allow the models that are within the top 1% of that category to predict agency-level consumption for each projection year. These consumption averages are aggregated, using growth estimates for the number of accounts in each member agency, to create a forecasted total demand in each future period. Thus, for each score category, we have a distribution of forecasted demand in each time period that is predicted by the top models in that score category. For each of these score category-prediction year combinations, we calculate the 10^{th}, 50^{th}, and 90^{th} percentiles of the aggregate forecast demand distribution.

The results in Table 5 confirm the trend we observed in the previous section, which explored 5-year prediction patterns across the best models for the different score categories. The types of models that score well for both of our in-sample score criteria, R^2 and AIC, predict increased demand over the full length of our prediction years. Across all the prediction years, the models that perform best by AIC criteria are less disperse, relative to the models that rank highest for R^2. However, the median aggregate forecasts for the AIC models are higher than the median for R^2 models in all prediction years. Despite these differences, the percent change in forecasted demand across the 25-year period is similar for these two scoring categories. The median path for the R^2 models increases by 8.4%, while the AIC models increase by 7.9% along the median prediction path.

The out-of-sample scoring methods, on the other hand, have lower median aggregate predictions in every period, relative to the in-sample score categories, and suggest a downward trend in future aggregate demand across our time horizon. The median prediction path for the RL-MSFE is essentially stable across time, with a 0.6% predicted reduction in demand over our prediction years. Forecasted demand decreases by 7.5% along the median path for AL-MSFE models. The models that performed best by aggregate forecast error undergoes the largest reduction in median predicted demand along the 25-year period, decreasing by 10.9%. We also see that, relative to the RL-MSFE and AL-MSFE models, the AFE models are more tightly distributed around the median.

For an alternative perspective, we graphically represent these results of forecast estimates in Figures 5 and 6. Here, we plot the average aggregate demand forecast in each prediction year for the top performing models in the relevant score category. We plot these averages in Figures 5 and 6, which includes a shaded region indicating two sample standard deviations above and below the mean, based on our sample of forecast models. Figure 5a and Figure 5b offer visual confirmation of both the upward trend in forecasted demand for the in-sample scoring methods and the tighter distribution of the AIC models, relative to the R^2 models. Similarly, we observe decreased average demand for all of the models that rank best in the out-of-sample score categories. As stated above, the models that were preferred by AFE standards demonstrate the largest reduction in average forecasted demand over the prediction time horizon. The distribution of these models is also the least disperse. Finally, we present the projection paths for both the R^2 and the AFE models on a single plot in Figure 7, allowing for a simpler visual comparison. This image illustrates the differences in both the average and the range of forecasted demand for these two scoring criteria.

7 Conclusion

The traditional methods of generating forecasted residential water consumption rely on both contestable structural assumptions and standard in-sample evaluation methods for model selection. These techniques are vulnerable to inaccurate demand forecasting, either due to misspecification or due to improper application of model performance evaluation. This paper provides evidence that using in-sample selection criteria to benchmark models for predicting future water demand may be misleading. We suggest an alternate procedure that employs evaluation methods which are consistent with and demonstrative of the water manager's and/or policy maker's objectives. Models are instead selected according to minimization of total forecast error along various levels of spatial aggregation to match forecast objectives. Using out-of-sample prediction ability for model selection criteria, not surprisingly, significantly improved precision of forecasted demand. We found that models that ranked highly for in-sample performance over-estimated for a 5-year prediction window by $10 - 25\%$, on average. Whereas, the top models from our out-of-sample criteria, came within 1% of the actual total consumption. This significant difference in 5-year prediction ability indicates the importance of incorporating forecasting goals in analyzing model performance. Notably, projections of future demand for the in-sample models indicate increasing aggregate water consumption over a 25-year period, which contrasts the downward trend predicted by the out-of-sample models. This non-trivial disparity is highly consequential in achieving effective water management for the state.

Allocation of California's scarce water resources among the various sectors of demand necessitates finding a best path forward through challenging choices. In particular, decisions about optimal investments, infrastructure, and policy conditions often entail analysis along extended time horizons. Under these circumstances, considering forecasts of residential water demand based on out-of-sample criteria is advisable when characterizing water resource management problems facing decision-makers. In this paper, we have shown that standard techniques of generating forecasts in SFR water demand produce projections that differ. We argue for a computationally-driven process that takes into account forecast objective, out-of-sample prediction ability, in casting projections of future residential water demand.

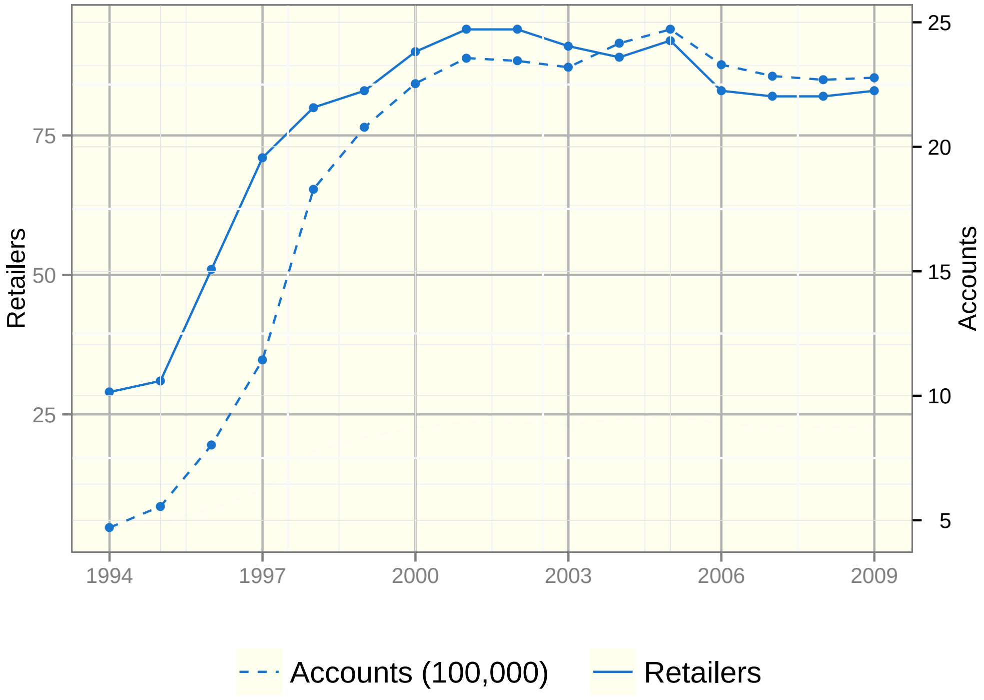


Figure 1: Availability of data in number of retailers and number of accounts across years. The solid line plots the number of retailers for which we have a data, out of a total 150 MWD retailers, in each year. The dashed line represents the total number of accounts (in hundred thousands) that is represented by the availability of retailer observations through time.

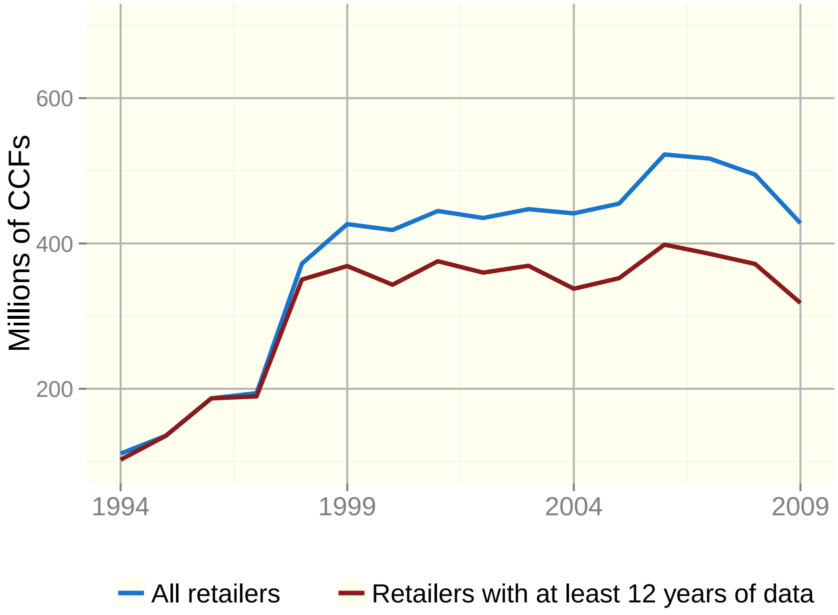
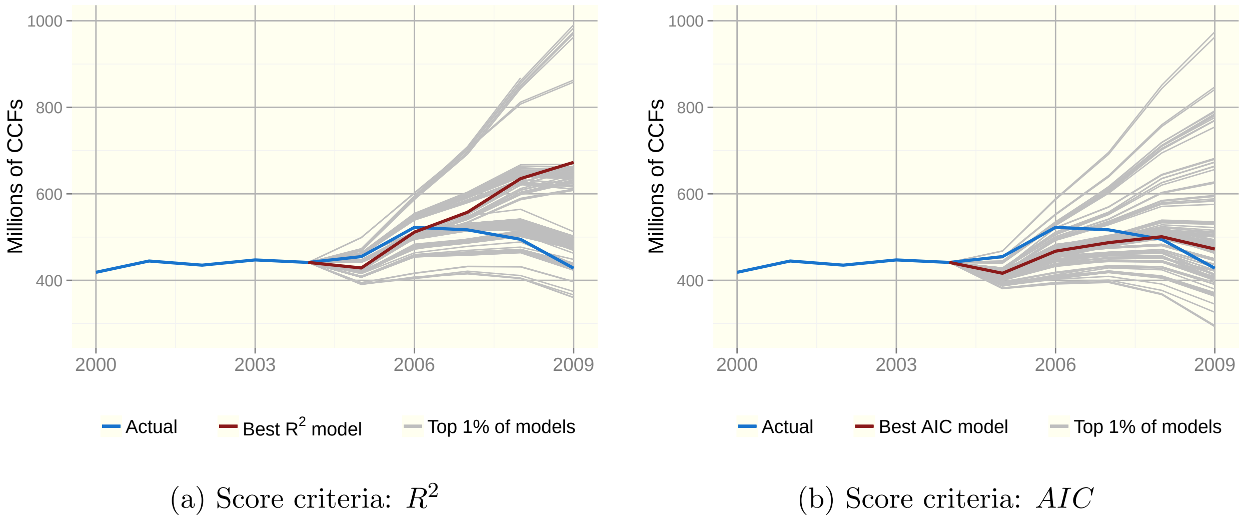


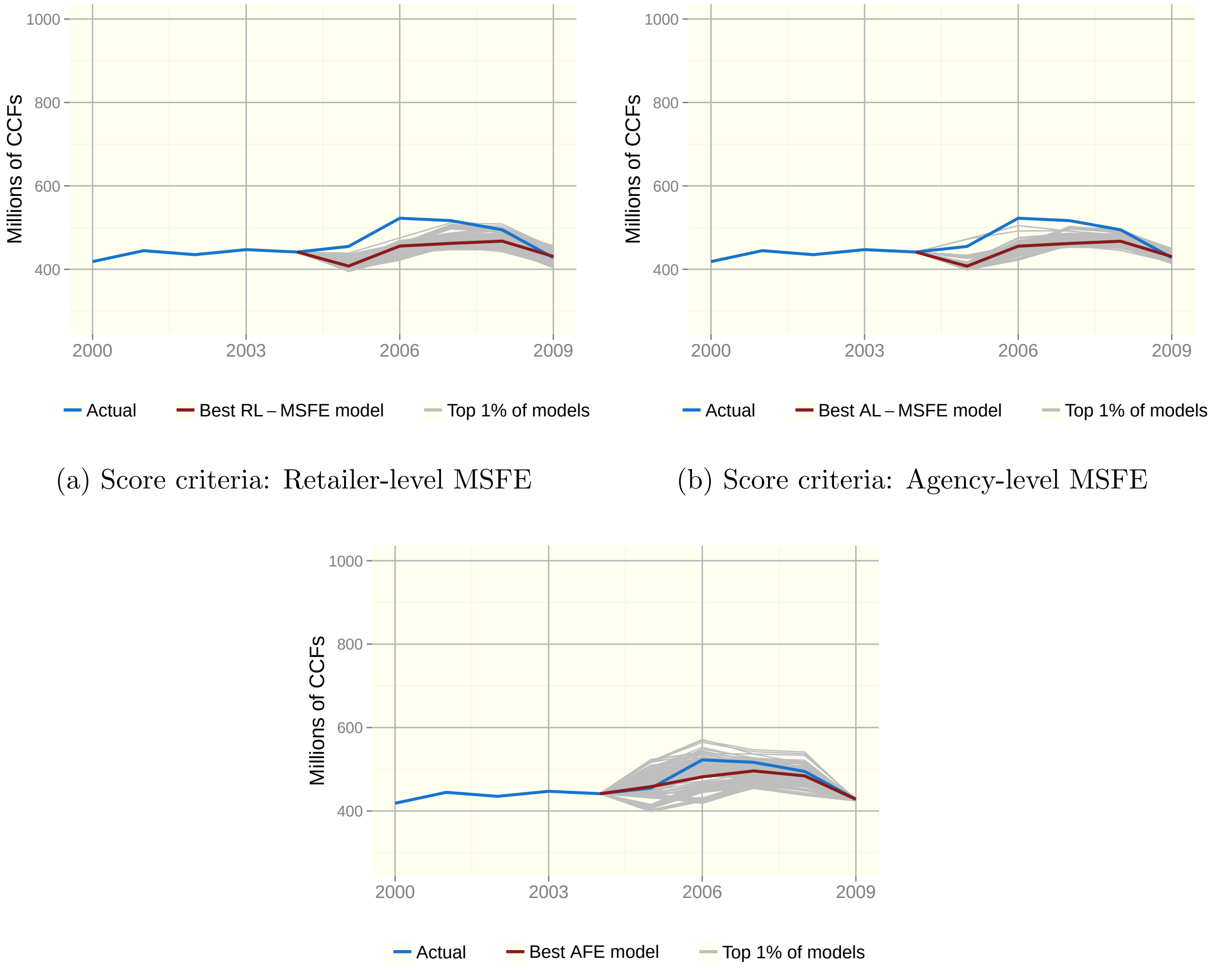
Figure 2: Aggregate annual demand for all retailers and aggregate annual demand for retailers with at least 12 years of data.



(a) Score criteria: R^2

(b) Score criteria: AIC

Figure 3: 2005-2009 predicted values for top 1% of models within each score criteria.

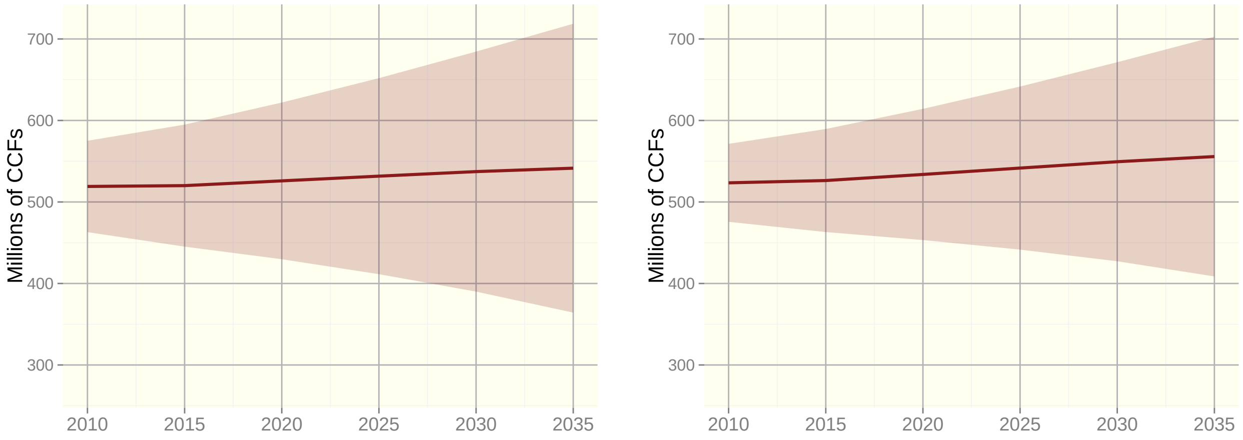


(a) Score criteria: Retailer-level MSFE

(b) Score criteria: Agency-level MSFE

(c) Score criteria: Aggregate forecast error

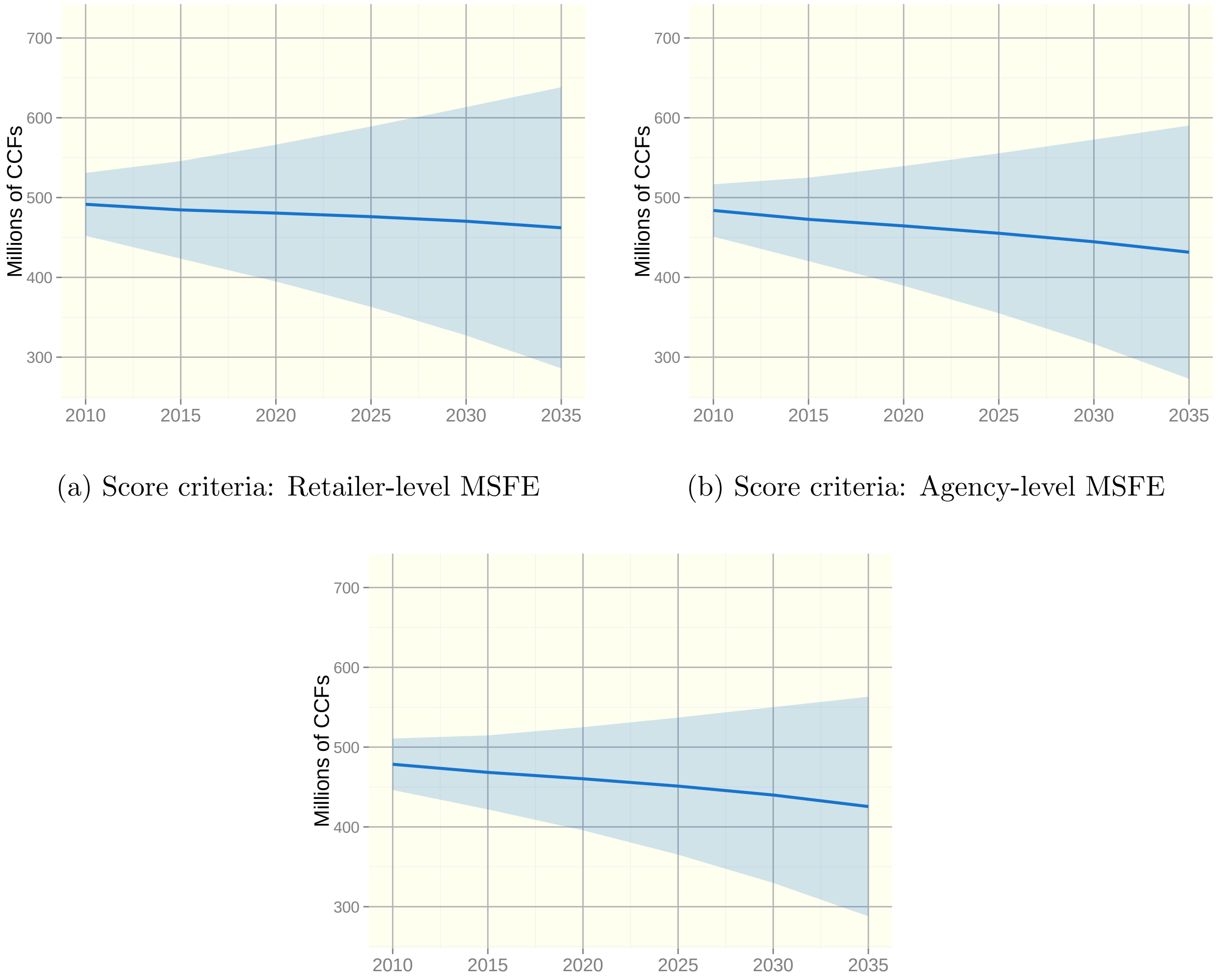
Figure 4: 2005-2009 predicted values for top 1% of models within each score criteria.



(a) Score criteria: R^2

(b) Score criteria: AIC

Figure 5: Predicted aggregate demand for top models within each score criteria.



(a) Score criteria: Retailer-level MSFE

(b) Score criteria: Agency-level MSFE

(c) Score criteria: Aggregate forecast error

Figure 6: Predicted aggregate demand for top models within each score criteria.

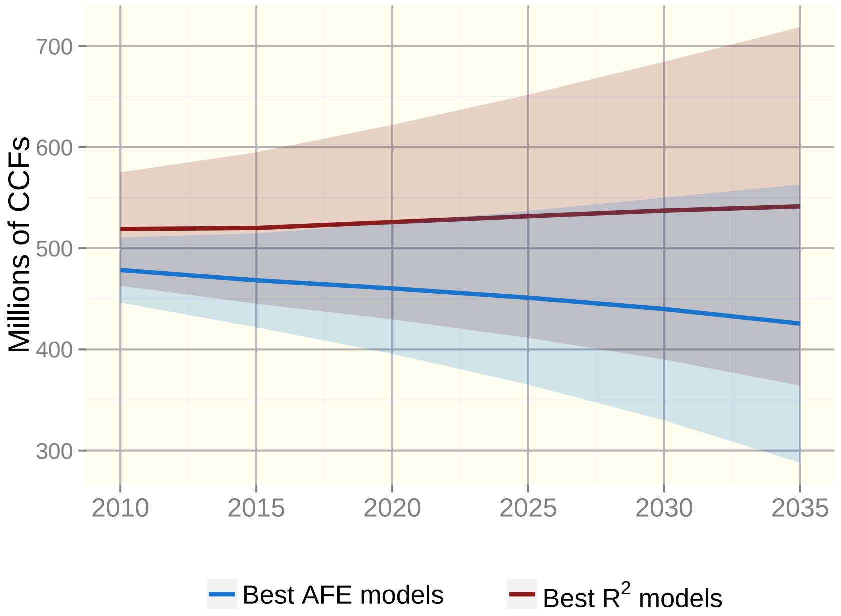


Figure 7: Forecasted aggregate demand of top models for two score categories.

Table 1: Retailer-level descriptive statistics.

Statistic	Mean	St. Dev.	Min	Max
Quantity (monthly CCFs)	22.04	8.83	5.68	69.19
Median tier rate	1.51	0.50	0.48	4.38
Total average cost	1.76	0.52	0.73	3.97
Precipitation (centimeters)	35.62	22.45	4.87	136.83
Average maximum temperature (celsius)	24.24	1.72	19.22	29.46
Average summer maximum temperature (celsius)	28.78	2.79	21.88	35.98
Median household income ($10,000)	6.19	1.90	2.75	12.27
Median lot size	9,796	8,327	4,957	60,548
Average household size (2000 census)	2.97	0.54	1.83	5.07
Number of accounts	24,655	51,160	523	484,042

Notes: Annual data from 1994-2009 , n=1225

Table 2: Number of retailers with various levels of data availability

	Number of retailers
All 16 years	15
Between 15 and 13 years	35
Between 12 and 10 years	20
Between 9 and 7 years	25
Less than 7 years	18

Table 3: Average values within each score category.

Score category	All models	Levels only	Logs only	No agency fixed effect	No lags
R^2	0.890	0.887	0.892	0.885	0.614
Adjusted R^2	0.885	0.882	0.886	0.884	0.599
AIC	2829	2834	2826	2827	5128
BIC	2871	2874	2869	2868	5280
RL-MSFE (billion CCFs)	456	484	434	443	1201
AL-MSFE (billion CCFs)	1815	2062	1615	1830	4501
Aggregate FE (thousand CCFs)	-3.51	-87.12	-17.17	63.89	131.90

Notes: Average for each score category across top one percent of models by model type. For each score category, all models are ranked for performance according to this criteria. We then calculate the average score criteria value of the top 1% of models as ranked by that criteria. Moving across columns for a particular score category, we report the average score criteria value for a subclass of models within that top 1%. RL-MSFE and AL-MSFE indicate retailer-level and agency-level mean square forecast error, respectively.

Table 4: Summary of average aggregate forecast error by score category and model type.

				No agency	
Score Category	All models	Levels only	Logs only	fixed effect	No lags
R^2	105.85	265.46	89.69	52.77	42.75
	[25%]	*[62%]*	*[21%]*	*[12%]*	*[10%]*
Adjusted R^2	84.8	185.51	82.58	49.19	43.89
	[20%]	*[43%]*	*[19%]*	*[11%]*	*[10%]*
AIC	64.29	113.62	49.12	50.76	46.47
	[15%]	*[27%]*	*[11%]*	*[12%]*	*[11%]*
BIC	42.76	53.25	40.84	43.52	50.48
	[10%]	*[12%]*	*[10%]*	*[10%]*	*[12%]*
RL-MSFE	-1.45	13.74	-8.1	-3.82	-9.93
	[-0.3%]	*[3%]*	*[-2%]*	*[-1%]*	*[-2%]*
AL-MSFE	5.74	10.28	2.57	5.36	-2.53
	[1%]	*[2%]*	*[1%]*	*[1%]*	*[-1%]*
Aggregate FE	-0.004	-0.09	-0.02	0.06	0.13
	[-0.001%]	*[-0.02%]*	*[-0.004%]*	*-0.020%]*	*[-0.03%]*

Notes: Average aggregate forecast error (in million CCFs) for each score category for the top one percent of models, with results also given for model subclasses. Percent deviation from actual aggregate demand, that each average AFE represents, is given in brackets below the average AFE. Moving across columns, for a particular score category, we see how the average AFE, and percent deviations, changes when the top 1% of models, by that score category, is restricted to particular specification characteristics RL-MSFE and AL-MSFE indicate retailer-level and agency-level mean square forecast error, respectively.

Table 5: Summary statistics for projected aggregate demand.

	2010	2015	2020	2025	2030	2035
R Squared						
10th percentile	483	470	458	446	429	408
50th percentile	511	515	525	536	545	554
90th percentile	561	572	587	612	638	664
Akaike Information Criteria						
10th percentile	495	490	486	482	475	462
50th percentile	522	525	533	541	552	563
90th percentile	561	573	588	611	637	664
Retailer-level mean square forecast error						
10th percentile	462	438	415	388	360	325
50th percentile	492	490	492	493	493	489
90th percentile	518	521	530	539	548	558
Agency-level mean square forecast error						
10th percentile	460	434	410	382	351	314
50th percentile	483	474	468	463	455	447
90th percentile	505	506	511	513	518	521
Aggregate forecast error						
10th percentile	458	438	416	392	362	329
50th percentile	478	466	460	453	443	426
90th percentile	500	499	505	510	515	519

Notes: Predicted total SFR demand (millions of CCFs) for the top 1% of models according to score category.

References

Arbués, Fernando, Marıa Ángeles Garcıa-Valiñas, and Roberto Martınez-Espiñeira. 2003. "Estimation of residential water demand: a state-of-the-art review." *The Journal of Socio-Economics* 32 (1):81–102.

Auffhammer, Maximilian and Richard T Carson. 2008. "Forecasting the path of China's CO 2 emissions using province-level information." *Journal of Environmental Economics and Management* 55 (3):229–247.

Auffhammer, Maximilian and Ralf Steinhauser. 2012. "Forecasting the path of US CO2 emissions using state-level information." *Review of Economics and Statistics* 94 (1):172–185.

Diffenbaugh, Noah S., Daniel L. Swain, and Danielle Touma. 2015. "Anthropogenic warming has increased drought risk in California." *Proceedings of the National Academy of Sciences* URL `http://www.pnas.org/content/early/2015/02/23/1422385112.abstract`.

Executive Office, State of California. 2015. "Executive Order B-29-15."

Herrera, Manuel, Luís Torgo, Joaquín Izquierdo, and Rafael Pérez-García. 2010. "Predictive models for forecasting hourly urban water demand." *Journal of hydrology* 387 (1):141–150.

Swain, Daniel L, Michael Tsiang, Matz Haugen, Deepti Singh, Allison Charland, Bala Rajaratnam, and Noah S Diffenbaugh. 2014. "The extraordinary California drought of 2013/2014: Character, context, and the role of climate change." *Bulletin of the American Meteorological Society* 95 (9):S3.

Zhou, Shuang Lin, Thomas Aquinas McMahon, Allan Walton, and Jane Lewis. 2000. "Forecasting daily urban water demand: a case study of Melbourne." *Journal of Hydrology* 236 (3):153–164.