Holistic Measures for Evaluating Prediction Models in Smart Grids

Saima Aman, *Student Member, IEEE,* Yogesh Simmhan, *Senior Member, IEEE,* and Viktor K. Prasanna, *Fellow, IEEE*

Abstract— The performance of prediction models is often based on "abstract metrics" that estimate the model's ability to limit residual errors between the observed and predicted values. However, meaningful evaluation and selection of prediction models for end-user domains requires holistic and application-sensitive performance measures. Inspired by energy consumption prediction models used in the emerging "big data" domain of Smart Power Grids, we propose a suite of performance measures to rationally compare models along the dimensions of scale independence, reliability, volatility and cost. We include both application independent and dependent measures, the latter parameterized to allow customization by domain experts to fit their scenario. While our measures are generalizable to other domains, we offer an empirical analysis using real energy use data for three Smart Grid applications: planning, customer education and demand response, which are relevant for energy sustainability. Our results underscore the value of the proposed measures to offer a deeper insight into models' behavior and their impact on real applications, which benefit both data mining researchers and practitioners.

Index Terms—Consumption Prediction, Performance Measures, Time series forecasting, Regression tree learning, Smart Grids, Energy Sustainability.

1 Introduction

There is a heightened emphasis on applying data mining and machine learning techniques to domains with societal impact, and, in the process, understanding the gaps in existing research [36]. One emerging community contending with a data explosion is Smart Power Grids. Advanced instrumentation and controls being deployed to upgrade aging power grid infrastructure is offering utilities unprecedented access to power data at fine spatial and temporal granularities, with near realtime availability [1] [30]. However, this data needs to be translated into actionable intelligence by means of novel modeling, mining and evaluation methods to ensure sustainable energy generation and supply [28].

One critical opportunity lies in *reliably, accurately and efficiently predicting electric energy consumption for individual premises*. The growing availability of energy consumption data (Kilowatt-Hour (kWh)) from 100,000's of customer smart meters at 15 mins sampling intervals [29] – orders of magnitude greater than traditional readings done once a month – offers a unique challenge in applying forecasting models and evaluating their efficacy for Smart Grid applications. The impact of such predictions includes strategic planning of renewable generation, power purchases from energy markets, daily planning to meet peak power loads, and engaging customers in energy savings programs [25], [37], all of which can enhance long term energy sustainability and security.

Historically, data at the feeder and sub-station level have been collected using SCADA systems. Hence contemporary load forecasting models exist at the coarser spatial granularity of total utility area using Bayesian modeling [10], Support Vector Machines [9], Artificial Neural Networks [21] and time series methods [23]. However, energy consumption prediction[2] for individual customers is less studied, both due to the lack of input data and the limited need for such predictions till recently with Smart Grids [15]. As a result, prediction models at the consumer level, with more intra-day and seasonal variability, and for innovative Smart Grid applications, which engage customers in sustainability, have not been well understood nor a *set of performance measures for their evaluation* identified.

The performance of prediction models are often based on "abstract metrics", detached from their meaningful evaluation for the end-user domain [36]. Common performance measures like Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) form the basis for selecting a suitable prediction model. *However, we suggest that these measures are in-*

- *S. Aman is with the Department of Computer Science, University of Southern California, Los Angeles, CA, 90089 E-mail: saman@usc.edu*
- *Y. Simmhan is with the Indian Institute of Science, Bangalore, India*
- *V. Prasanna is with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA, 90089*

1. Smart Meter deployments continue to rise, US Energy Info. Admin., 2012 www.eia.gov/todayinenergy/detail.cfm?id=8590

2. In this article, we address energy consumption prediction, which deals with average energy over an interval (i.e., kWh). This is different from demand (or load) prediction [2], [33] that deals with predicting instantaneous power (kW).

sufficient to evaluate prediction models for emerging Smart Grid applications. This gap, discussed below, is relevant to many applied domains beyond just Smart Grids.

(1) The impact of under- and over- predictions can be asymmetric on the domain, and measures like RMSE are insensitive to *prediction bias*. For e.g., under-prediction of consumption forecasts is more deleterious to Smart Grid applications that respond to peak demand. **(2)** *Scale-dependent* metrics are unsuitable for comparing prediction models applied to different customer sizes. **(3)** The focus on the magnitude of errors overlooks the *frequency* with which a model outperforms a baseline model or predicts within an error tolerance. Reliable prediction is key for certain domain applications. **(4)** *Volatility* is a related factor that is ignored in common measures, wherein a less volatile prediction model performs *consistently* better than a baseline model. **(5)** Lastly, given the "Big Data" consequences of emerging applications, the *cost of collecting data and running models* cannot be disregarded. The extra cost for improved accuracy from a model may be impractical at large scales in a Smart Grid with millions of customers [18], or the latency for a prediction can make it unusable for operational decisions. These gaps highlight the need for holistic performance measures to meaningfully evaluate and compare prediction models by domain practitioners.

Contributions. Specifically, we make the following novel contributions. (1) We propose a suite of performance measures for evaluating prediction models in Smart Grids, defined along three dimensions: scale in/dependence, reliability and cost (§3). These include two existing measures and eight innovative ones (§4, §5), and also encompass parameterized measures that can be customized for the domain. (2) We analyze the usefulness of these concrete measures by evaluating ARIMA and regression tree prediction models (§6) applied to three Smart Grid applications (§7) in the Los Angeles Smart Grid Project [3].

Significance. In this article, we offer meaningful *performance measures* to evaluate predictive models along dimensions that go beyond just the *magnitude of errors*, and explore bias, reliability, volatility and cost [36]. Not all our measures are novel and some extend from other disciplines – this offers completeness and also gives a firm statistical grounding. Our novel application dependent measures with parameterized coefficients set by domain experts allow apples-to-apples comparison that is meaningful for that scenario [12]. A model that seems good using common error metrics may behave poorly or prove inadequate for a given application; this intuition is validated by our analysis. All our measures are reusable by other domains, though they are inspired by and evaluated for the emerging Smart Grid domain.

As Smart Grid data becomes widely available, data mining and machine learning research can provide immense societal benefits to this under-served domain [28]. Our study based on real Smart Grid data collected over 3 years [4] is among the *first of its kind* in defining holistic measures and evaluating candidate consumption models for emerging microgrid and utility applications. Our analysis of the measures underscores their key ability to offer: deeper insight into models' behavior that can help improve their performance, better understanding of prediction impact on real applications, intelligent cost-benefit trade-offs between models, and a comprehensive, yet accessible, goodness of fit for picking the *right* model. Our work offers a common frame of reference for future researchers and practitioners, while also exposing gaps in existing predictive research for this new domain.

2 Related Work

The performance evaluation of predictive models often involves a single dimension, such as an error measure, which is simple to interpret and compare, but does not necessarily probe all aspects of a model's performance or its goodness for a given application [36]. Davydenko and Fildes highlight that existing error measures are not suited for evaluating forecast adjustments and may lead to false conclusions [12]. They propose a new metric based on aggregating performance ratios across time series for fair treatment of over- and under-forecasting. Recently, Prati, et al. [27] emphasized the importance of treating predictive performance as a multi-dimensional problem for a more reliable evaluation of the trade-offs between various aspects. Further, [35] has identified the need for cost-benefit measures to help capture the performance of a prediction model by a single profit-maximization metric which can be easily interpreted by practitioners. Armstrong and Collopy [5] have highlighted the importance of scale, along with measures like cost, sensitivity, reliability, understandability, and relationship for decision making using forecasts. Other studies [32] also go beyond standard error measures to include dimensions of sensitivity and specificity. Our effort is in a similar vein. We propose a holistic set of measures along multiple dimensions to assist domain users in intelligent model selection for their application, with empirical validation for emerging Smart Grid domain.

Existing approaches for consumption prediction include our and other prior work on regression tree [3], [26], time series models [4], artificial neural networks and expert systems [21], [22], [23]. In practice, utilities use simpler averaging models based on recent consumption [1], [11]. In this spirit, our baseline models for comparative evaluation consider Time of the Week (ToW) and Day of the Week (DoW) averages.

3. Los Angeles Department of Water and Power: Smart Grid Regional Demonstration, US Department of Energy, 2010

4. "Where's the Data in the Big Data Wave?", Gerhard Weikum, SIGMOD Blog, Mar 2013. wp.sigmod.org/?p=786

As elsewhere, Smart Grid literature often evaluates predictive model performance in terms of the magnitude of errors between the observed and predicted values [8], [19]. Common statistical measures for time series forecasting [5] are *Mean Absolute Error (MAE)*, the *Root Mean Square Error (RMSE)* and the *Mean Absolute Percent Error (MAPE)*, the latter given by:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|p_i - o_i|}{o_i} \quad (1)$$

where o_i is the observed value at interval i, p_i is the model predicted value, and n is the number of intervals for which the predictions are made. *Mean Error Relative (MER)* to the mean of the observed, has also been used to avoid the effects of observed values close to zero [23]. RMSE values normalized by the mean of observed values, called the *coefficient of variation of the RMSE (CVRMSE)*, is also used [3], [13]:

$$CVRMSE = \frac{1}{\bar{o}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2} \quad (2)$$

where $\bar{o}$ is the mean of the n observed values, and p_i, o_i and n are as before. While these measures offer a necessary statistical standard of model performance, they by themselves are inadequate due to the reasons listed before, viz., their inability to address prediction-bias, reliability, scale independence and cost of building models and making predictions.

Some researchers have proposed application-specific metrics. Mathieu, et al. [24] define metrics related to Demand-Response. The Demand Shed Variability Metric (SVM) and Peak Demand Variability Metric (PVM) help reduce over-fitting and extrapolation errors that increase error variance or introduce prediction bias. Our application-*dependent* (rather than -*specific*) measures are defined more broadly, with measure *parameters* that can be tuned for diverse applications that span even beyond Smart Grids.

Relative measures help compare a prediction model with a baseline model [5]. *Percent Better* gives the fraction of forecasts by a model that are more accurate than a random walk model. This is a unit-free measure that is also immune to outliers present in the series, by discarding information about the amount of change. The *Relative Absolute Error (RAE)*, calculated as a ratio of forecast error for a model to the corresponding error for the random walk, is simple to interpret and communicate to the domain users. The prediction horizon also has an impact on model performance. *Cumulative RAE* [5] is defined as the ratio of the arithmetic sum of the absolute error for the proposed model over the forecast horizon and the corresponding error for the random walk model. Relative and horizon metrics have been used less often for smart grid prediction models.

Prediction error metrics have been categorized into scale-dependent measures, percentage errors, relative errors and scaled errors [16]. RMSE and MAE are *scale-dependent* and applicable to datasets with similar values. Scale-independent *Percentage errors* like MAPE and CVRMSE can be used to compare performance on datasets with different magnitudes. *Relative measures* are determined by dividing model errors by the errors of a baseline model, while *scaled errors* remove the scale of data by normalizing the errors with the errors obtained from a baseline prediction method.

In summary, standard statistical measures of performance for predictive models may not be adequate or meaningful for domain-specific applications, while narrowly defined measures for a single application are not reusable or comparable across applications. This gap is particularly felt in the novel domain of Smart Grids. Our work is an attempt to address this deficiency by introducing a suite of performance measures along several dimensions, while also leveraging existing measures where appropriate.

3 Performance Measure Dimensions

Performance measures that complement standard statistical error measures for evaluating prediction models fall along several dimensions that we discuss here.

Application In/dependent: Application independent measures are specified without knowledge of how predictions from the model are being used. These do not have any specific dependencies on the usage scenario and can be used as a uniform measure of comparison across different candidate models. Application dependent measures incorporate *parameters* that are determined by specific usage scenarios of the prediction model. The measure formulation itself is generic but requires users to set values of parameters (e.g., acceptable error thresholds) for the application. These allow a nuanced evaluation of prediction models that is customized for the application in concern.

Scale-Independent Errors: In defining error metrics, the residual errors are measured as a difference between the observed and predicted values. So, if o_i is the i^{th} observed value and p_i is the i^{th} predicted value, then the *scale-dependent* residual or prediction error, $e_i = p_i - o_i$. MAE and RMSE are based on residual errors, and suffer from being highly dependent on the range of observed values. Scale-independent errors, on the other hand, are usually normalized against the observed value and hence better suited for comparing model performance across different magnitudes of observations.

Reliability: Reliability offers an estimate of how consistently a model produces similar results. This dimension is important to understand how well a model will perform on a yet unseen data that the system will encounter in future, relative to the data used while testing. A more reliable model provides its users

with more confidence in its use. Most commonly used measures fail to consider the frequency of acceptable model performance over a period of time, which we address through measures we introduce.

Cost: Developing a prediction model has a cost associated with it in terms of effort and time for data collection, training and using the models. The number of times a trained model can be reused is also a factor. Data cost is a particularly important consideration in this age of "Big Data" since quality checking, maintaining and storing large featuresets can be untenable. Compute costs can even be intractable when prediction models are used millions of times within short periods.

4 Application Independent Measures

Several standard statistical measures with well understood theoretical properties fall in the category of application independent measures. For completeness, we recognize two relevant, existing scale-independent measures, MAPE [19], [23] and CVRMSE [3], [13]. More importantly, we introduce novel application independent measures, along the reliability and cost dimensions, and their properties.

4.1 Mean Absolute Percentage Error (MAPE)

This is a variant of MAE, normalized by the observed value[5] at each interval (1), thus providing *scale independence*. It is simple to interpret and commonly used for evaluating predictions in energy and related domains [8], [9], [17], [19], [23].

4.2 Coefficient of Variation of RMSE (CVRMSE)

It is the normalized version of the common RMSE measure, that divides it by the average [6] of the observed values (2) to offer *scale independence*. This is an unbiased estimator that incorporates both the prediction model bias and its variance, and gives a unitless percentage error measure. CVRMSE is sensitive to infrequent large errors due to the squared term.

4.3 Relative Improvement (RIM)

We propose RIM as a *relative measure* for *reliability* that is estimated as the frequency of predictions by a candidate model that are better than a baseline model. RIM is a simple, unitless measure that complements error measures in cases where being accurate more often than a baseline is useful, and occasional large errors relative to the baseline are acceptable.

$$RIM = \frac{1}{n}\sum_{i=1}^{n} C(p_i, o_i, b_i) \tag{3}$$

where o_i, p_i and b_i are the observed, model predicted and baseline predicted values for interval i, and $C(p_i, o_i)$ is a count function defined as:

$$C(p_i, o_i, b_i) = \begin{cases} 1, & \text{if } |p_i - o_i| < |b_i - o_i| \\ 0, & \text{if } |p_i - o_i| = |b_i - o_i| \\ -1, & \text{if } |p_i - o_i| > |b_i - o_i| \end{cases} \tag{4}$$

4.4 Volatility Adjusted Benefit (VAB)

VAB is another measure for *reliability* that captures how consistently a candidate model outperforms a baseline model by normalizing the model's error improvements over the baseline by the standard deviation of these improvements. Inspired by the *sharpe ratio*, this *relative measure* offers a "risk adjusted" scale-independent error value. The numerator captures the relative improvement of the candidate model's MAPE over the baseline's (the benefit). If these error improvements[7] are consistent across i, then their standard deviation would be low (the volatility) and the VAB high. But, with high volatility, the benefits would reduce reflecting a lack of consistent improvements.

$$VAB = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{|b_i - o_i|}{o_i} - \frac{|p_i - o_i|}{o_i}\right)}{\sigma\left(\frac{|b_i - o_i|}{o_i} - \frac{|p_i - o_i|}{o_i}\right)} \tag{5}$$

where o_i, p_i and b_i are the observed, model predicted and baseline predicted values for interval i.

4.5 Computation Cost (CC)

The *cost* for training and predicting using a model can prove important when it is used either at large scales and/or in realtime applications that are sensitive to prediction latency. CC is defined in seconds as the sum of the wallclock time required to train a model, CC_t, and the wallclock time required to make predictions using the model, CC_p, for a given prediction duration with a certain horizon. Thus, $CC = CC_t + CC_p$.

4.6 Data collection Cost (CD)

Rather than examine the raw size of data used for training or predicting using a model, a more useful measure is the effort required to acquire and assemble the data. Size can be managed through cheap storage but collecting the necessary data often requires human and organizational effort. We propose a *scale-dependent* measure of data *cost* defined in terms of the number of *unique values* of features involved in a prediction model. CD is defined for a particular training and prediction *duration* as the sum of n_s, the number of static (time-invariant) features that require a one-time collection effort, and n_d, the number of dynamic features that need periodic acquisition.

5. MAPE is not defined if there are zero values in the input, which is rare as energy consumption (kwh) values are generally non-zero due to always present base consumption (unless there is a blackout), and can be ensured by data pre-processing.

6. CVRMSE is not defined if this average is zero, which is rare as energy consumption (kwh) values are generally positive (unless there is net-metering), and can be ensured by data pre-processing.

7. The error improvements offered by a given model over the baseline model are expected to have normal distribution for VAB to be meaningful.

$$CD = \sum_{i=1}^{n_s}[s_i] + \sum_{i=1}^{n_d}[d_i] \quad (6)$$

where $[s_i]$ and $[d_i]$ are the counts of the unique values for the feature s_i and d_i respectively.

5 Application Dependent Measures

Unlike the previous measures, application dependent performance measures are *parameterized* to suit specific usage scenarios and can be customized by domain experts to fit their needs. The novel measures we propose here are themselves not narrowly defined for a single application (though they are motivated by the needs observed in the smart grid domain). Rather, they are generalized through the use of coefficients that are themselves application specific. We group them along the dimensions that we introduced earlier.

5.1 Domain Bias Percentage Error (DBPE)

We propose DBPE as a signed *percentage error* measure that offers *scale independence*. It indicates if the predictions are positively or negatively biased compared to the observed values, which is important when over- or under-prediction errors, relative to observed, have a non-uniform impact on the application. We define DBPE as an asymmetric loss function based on the sign bias. Granger's linlin function [14] is suitable for this as it is linear on both sides of the origin but with different slopes on each side. The asymmetric slopes allow different penalties for positive/negative errors.

$$DBPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathcal{L}(p_i, o_i)}{o_i} \quad (7)$$

where $\mathcal{L}(p_i, o_i)$ is the linlin loss function defined as:

$$\mathcal{L}(p_i, o_i) = \begin{cases} \alpha \cdot |p_i - o_i|, & \text{if } p_i > o_i \\ 0, & \text{if } p_i = o_i \\ \beta \cdot |p_i - o_i|, & \text{if } p_i < o_i \end{cases} \quad (8)$$

where o_i and p_i are the observed and model predicted values for the interval i, and α and β are penalty parameters associated with over- and under- prediction, respectively. α and β are configured for specific application and the ratio α/β measures the relative cost of over-prediction to under-prediction for that application [34]. Further, we introduce a constraint that $\alpha + \beta = 2$ to provide DBPE the interesting property of reducing to MAPE when $\alpha = \beta = 1$.

5.2 Reliability Threshold Estimate (REL)

Often, applications may care less about the absolute errors of a model's predictions and prefer an estimate of how frequently the errors fall within a set threshold that the application can withstand. We define REL as the frequency of prediction errors that are less than an application determined error threshold, e_t.

$$REL = \frac{1}{n}\sum_{i=1}^{n} C(p_i, o_i) \quad (9)$$

where o_i and p_i are the observed and the model predicted values for the interval i, and $C(p_i, o_i)$ is a count function defined as:

$$C(p_i, o_i) = \begin{cases} 1, & \text{if } \frac{|p_i - o_i|}{o_i} < e_t \\ 0, & \text{if } \frac{|p_i - o_i|}{o_i} = e_t \\ -1, & \text{if } \frac{|p_i - o_i|}{o_i} > e_t \end{cases} \quad (10)$$

5.3 Total Compute Cost (TCC)

In the context of an application, it is meaningful to supplement the data and compute costs (CD and CC) with an estimate of the total running cost of using a model *for a duration of interest* specific to that application. We define the parameters:

- τ, the number of times a model is trained within the duration,
- π, the number of times a model makes predictions with a given horizon, in that duration.

These parameters are not just application specific but also vary by the candidate model, based on how frequently it needs to be trained and its effective prediction horizon. We define the total training cost in seconds for a prediction duration based on τ and π, and the unit costs for training and prediction using the model, CC_t and CC_p, introduced in § 4.5:

$$TCC = CC_t \cdot \tau + CC_p \cdot \pi \quad (11)$$

5.4 Cost-Benefit Measure (CBM)

Rather than treat cost in a vacuum, it is worthwhile to consider the cost for a model relative to the gains it provides. CBM compares candidate models having different error measures and costs to evaluate which provides a high reward for a unit compute cost spent.

$$CBM = \frac{(1 - DBPE)}{TCC} \quad (12)$$

The numerator is an estimate of the accuracy (one minus error measure) while the denominator is the compute cost. We use DBPE as the error measure and TCC as the cost, but these can be replaced by other application dependent error measures (e.g., CVRMSE, MAPE) and costs (e.g., CD, CC_p). A model with high accuracy but prohibitive cost may be unsuitable.

6 Candidate Prediction Models

The candidate models for evaluation of the proposed measures were selected based on our prior study [3] as well as existing literature discussed in § 2.

Time Series Model: A time series (TS) model predicts the future values of a variable based on its previous observations. The ARIMA (Autoregressive Integrated Moving Average) model is a commonly used TS prediction model. It is defined in terms of the number (d) of times a time series needs to be differenced to make it stationary; the autoregressive order

(p) that captures the number of past values; and the moving average order (q) that captures the number of past white noise error terms. These parameters are determined using autocorrelation and partial autocorrelation functions, using the Box-Jenkins test [6].

ARIMA is simple to use as it does not require knowledge of the underlying domain [20]. However, estimating the model parameters, $d, p,$ and q, requires human examination of the partial correlogram of the time series, though some automated functions perform a partial parameter sweep to select these values.

Regression Tree Model: A regression tree (RT) model [7] is a kind of decision tree that recursively partitions the data space into smaller regions, until a constant value or a linear regression model can be fit for the smallest partition.

Our earlier work on an RT model for campus microgrid consumption prediction identified several advantages [3]. It's flowchart style tree structure helps interpret the impact of different features on consumption. Making predictions on a trained model is fast though collecting feature data and training the model can be costly. It can be used to make predictions far into the future if the feature values are available.

7 Experimental Setup

We validate the efficacy of our proposed performance measures for real world applications. The USC campus microgrid [31] is a testbed for the DOE-sponsored Los Angeles Smart Grid Project. ARIMA and Regression Tree prediction models are used to predict energy consumption at 24-hour and 15-min granularities, for the entire campus and for 35 individual buildings. Here, we consider the campus and four representative buildings: *DPT*, a small department with teaching and office space; *RES*, a suite of residential dormitories with decentralized control of cooling and appliance power loads; *OFF*, hosting administrative offices and telepresence lab; and *ACD*, a large academic teaching building. These buildings were considered after several pilot studies to provide diversity in terms of floor size, age, end use, types of occupants, and net electricity consumption.

7.1 Datasets

Electricity Consumption Data [8]**:** We used 15-min granularity electricity consumption data collected by the USC Facility Management Services between 2008 to 2010 (Table 1). These gave $3 \times 365 \times 96$ or $\sim$100K samples per building. We linearly interpolated missing values (<3% of samples) and aggregated 15-min data in each day to get the 24-hour granularity values ($\sim$1K samples per building). Observations from 2008 and 2009 were used for training the models while the predictions were evaluated against the out-of-sample observed values for 2010 [9].

TABLE 1: **Electricity consumption dataset.** Summary statistics of the campus microgrid consumption data for training years 2008-2009, and testing year 2010, at different spatial and temporal granularities.

Entity	Mean *(kWh)*		Std. Deviation *(kWh)*	
	Training	Testing	Training	Testing
Campus				
24-hour data	462,970	440,803	52,956	43,454
15-min data	4,823	4,377	809	770
DPT				
24-hour data	405.64	405.56	112.39	108.14
15-min data	4.23	4.16	1.93	1.98
RES				
24-hour data	4,220.30	3,670.56	1,809.00	1,460.08
15-min data	43.97	37.79	22.36	17.93
OFF				
24-hour data	2,938.90	2,790.70	591.97	549.37
15-min data	30.66	28.42	13.05	10.03
ACD				
24-hour data	4,466.40	4,055.85	640.92	552.64
15-min data	46.65	41.30	14.08	13.09

Weather Data [10]**:** We collected historical hourly average and maximum temperature data curated by NOAA for Los Angeles/USC Campus for 2008–2010. These values were linearly interpolated to get 15-min values. We also collected daily maximum temperatures that were used for the 24-hour granularity models.

Schedule Data [11]**:** We gathered campus information related to the semester periods, working days and holidays from USC's Academic Calendar.

7.2 Model Configurations

Regression Tree (RT) Models: For 24-hour (granularity) predictions, we used five features for the RT model: Day of the Week (Sun-Sat), Semester (Fall, Spring, Summer), Maximum and Average Temperatures, and a Holiday/Working day flag. For the 15-min (granularity) predictions, we used five features: Day of the Week, Time of Day (1–96, representing the 15-min slots in a day), Semester, temperature, and Holiday/Working day flag. The RT model was trained once using *MATLAB's* `classregtree` function [7] to find an optimally pruned tree.

ARIMA Time Series (TS) Models: For 24-hour predictions, the ARIMA models are *retrained* and used to make predictions every week for four different prediction horizons: 1-week, 2-week, 3-week, and 4-week ahead. Unlike RT, the performance of time series models differ by the prediction horizon. We use a moving window over the past 2 years for training

8. The electric consumption datasets used in this article are available upon request for academic use.

9. The 24-hour data was available only till Nov 2010 at the time of experiments, and hence 24-hour models are tested for a 11 month period. The 15-min models span the entire 12 months.

10. NOAA Quality Controlled Local Climatological Data, cdo.ncdc.noaa.gov/qclcd/

11. USC Academic Calendar, academics.usc.edu/calendar/

these models with $(p, d, q) = (7, 1, 7)$, equivalent to a 7 day lag, selected after examining several variations. For 15-min predictions, we *retrain* models and predict every 2 hours for three different horizons: 2-hour, 6-hour, and 24-hour ahead. We use a moving window over the past 8 weeks for training, with $(p, d, q) = (8, 1, 8)$, equivalent to a 2 hour lag. We used the `arima` function in the *R forecast package* [17] for constructing the time series models. This function used conditional sum of squares (CSS) as the fitting method.

Baseline Models: For 24-hour predictions, we selected the Day of Week mean (DoW) as the baseline, defined for each day of the week as the kWh value for that day averaged over the training period (i.e., 7 values from averaging over 2008 and 2009). DoW was chosen over Day of Year (DoY) and Annual Means since it consistently out-performed them. For 15-min predictions, we selected the Time of the Week mean (ToW) as the baseline, defined for each 15-min in a week as the kWh value for that interval averaged over the training period (i.e., 7×96 values from averaging over 2008 and 2009). Here too, ToW out-performed Time of the Year (ToY) and Annual Means.

7.3 Smart Grid Applications

We introduce three applications, used within the USC microgrid, to evaluate our proposed measures.

Planning: Planning capital infrastructure such as building remodeling and power system upgrades for energy efficiency trades-off investment against electric power savings. Medium to long term electricity consumption predictions at coarse (24-hour) granularity for campus and individual buildings help this decision making. Such models are run six times in a year.

Customer Education: Educating power customers on their energy usage can enhance their participation in energy sustainability by curtailing demand and meeting monthly budgets [28]. One form of education is through giving consumption forecasts to customers in a building on web and mobile apps [12]. Building-level predictions at 24-hour and 15-min granularities are made during the day (6AM-10PM).

Demand Response: Demand Response (DR) optimization is a critical technique for achieving energy sustainability enabled by smart grids. In DR, customers are encouraged to curtail consumption, *on-demand*, to reduce the chance of black-outs when peak power usage periods are anticipated by the utility [24]. Historically, these high peaks occur between 1-5PM on weekdays [13], and predictions during these periods over the short time horizon at 15-min granularity are vital for utilities to decide when to initiate curtailment requests from customers or change their pricing. Often, the predictions are before, at the beginning of, and during the high peak period.

12. USC SmartGrid Portal, smartgrid.usc.edu

13. DWP TOU Pricing, bp.ladwp.com/energycredit/ energycredit.htm

8 Analysis of Independent Measures

We first examine the use and value of the six application independent measures (§ 4) to evaluate the candidate models for predicting campus and building consumption at coarse and fine time granularities.

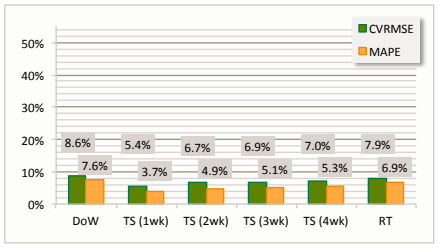
8.1 24-hour Campus Predictions

Fig. 1a presents the CVRMSE and MAPE measures for the DoW baseline, RT, and TS models, the latter for four different horizons, for campus 24-hour predictions. By these measures, TS models at different horizons offer higher accuracy than the RT and DoW models. This is understandable, given the noticeable difference in mean and standard deviations (Table 1) between the training and test periods. TS incrementally uses *more recent data as a moving window*, while RT and DoW model are only trained on the two years' test data. Also, the errors for TS deteriorate as the prediction horizon increases. This is a consequence of their dependence on recent lag values, making them suited only for *near-term predictions*. RT models are *independent of prediction horizons* (assuming future feature values are known), and therefore preferable for predictions with long horizons. The DoW errors are *marginally* higher than RT. This is quickly evident using our relative improvement (RIM) measure (Fig. 2a), that reports an improvement of 2.5% for RT and 58.39% for TS (1wk) over the baseline. However, when volatility is accounted for, this margin over the DoW increases to a VAB of 11.45% and 74.42% for RT and TS (1wk), making them *much more dependable*.

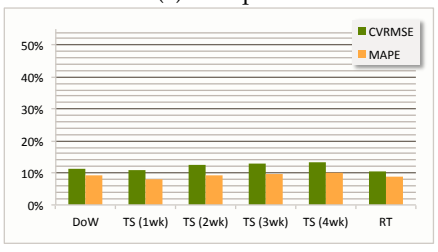
8.2 24-hour Building Predictions

The CVRMSE and MAPE measures for DPT (Fig. 1b) diverge in their ranking of the RT and TS models; RT is best based on CVRMSE while TS (1wk) is best on MAPE. This divergence highlights the value of having different error measures. In CVRMSE, residual errors are squared and thus *large errors are magnified* more than in MAPE. Our RIM measure offers another perspective as a relative measure independent of error values (Fig. 2b). TS (1wk) *is clearly more favorable* than RT, performing better than the baseline in 50% of predictions (RIM ≈ 0) compared to RT (RIM=-19.88%). When accounting for volatility in VAB, TS (1wk) outperforms the DoW (VAB=17.62%) and even RT exhibits lesser relative volatility (VAB=4.35%). These demonstrate why multiple measures offer a more holistic view of the model performance.

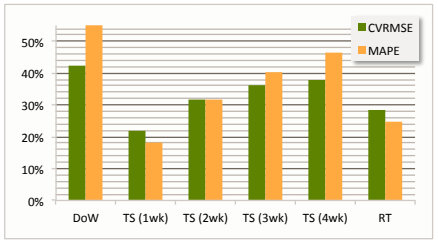
RES has 100's of residential suites with independent power controls, and hence higher consumption variability. This accounts for the higher errors in predictions across models (Fig. 1c). Further, the building is *unoccupied during summer* and vacation periods. Hence, it is unsurprising to see DoW perform particularly worse. (We verified the impact of summer by comparing DoW with DoY. DoY did perform better,



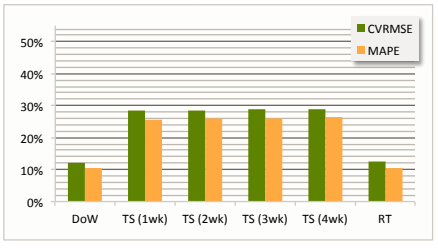
(a) Campus



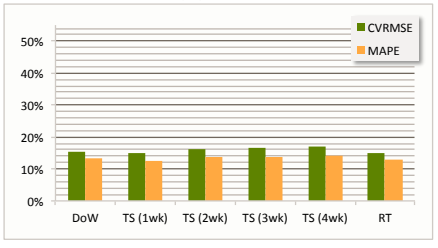
(b) DPT



(c) RES

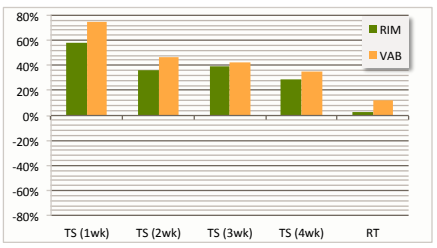


(d) OFF

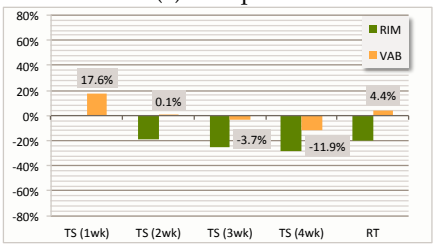


(e) ACD

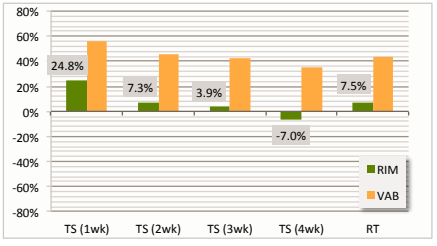
Fig. 1: ***CVRMSE*** and ***MAPE*** values for ***24-hour predictions*** for campus and four buildings. Lower values are better. Day of Week (DoW) baseline, ARIMA Time Series (TS) with 1, 2, 3 & 4-week prediction horizons, and Regression Tree (RT) models are on X-axis. Campus has the smallest errors, RES residential building the largest, and, except for OFF, TS and RT outperform the baseline.



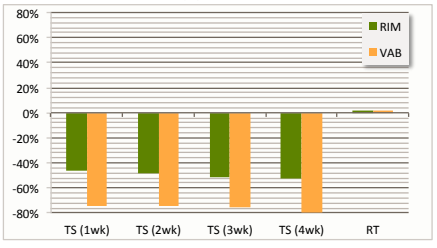
(a) Campus



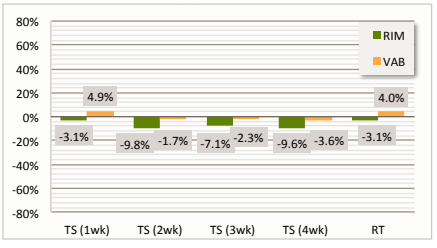
(b) DPT



(c) RES

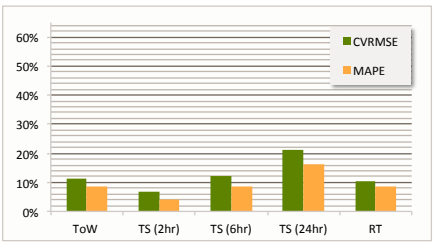


(d) OFF

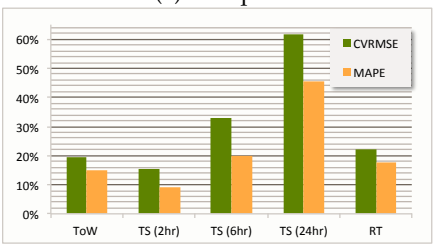


(e) ACD

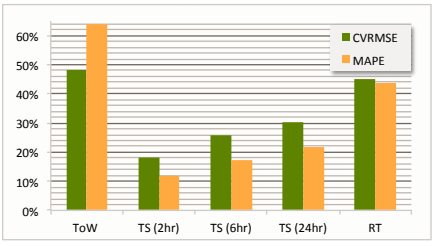
Fig. 2: Relative Improvement (***RIM***) and Volatility-Adjusted Benefit (***VAB***) values for ***24-hour predictions*** for campus and four buildings. Higher values indicate better performance relative to DoW baseline; zero value means performance similar to baseline. DoW is more volatile for RES due to summer vacation. VAB for RT and TS are high, showing resilience.



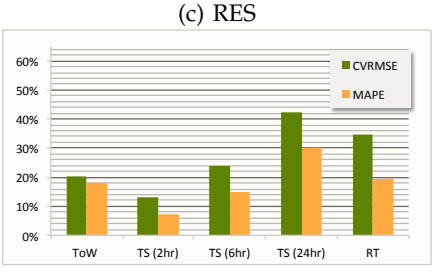
(a) Campus



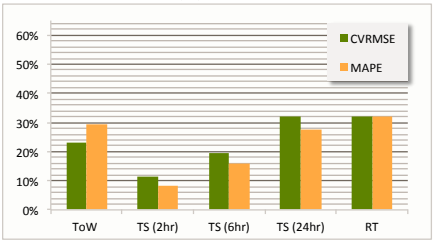
(b) DPT



(c) RES



(d) OFF



(e) ACD

Fig. 3: ***CVRMSE*** and ***MAPE*** values for ***15-min predictions*** for campus and four buildings. Lower errors indicate better model performance. Time of Week baseline (ToW), ARIMA Time Series (TS) at 2, 6 & 24-hour horizons, and Regression Tree (RT) models are shown. Errors usually increase as the prediction horizon is increased. RT is independent of prediction horizon.

but for consistency, we retain DoW as the baseline.) RT has lower errors than DoW as it captures *schedule-related features like holidays and summer semester*. However, the test data for RES has a smaller mean than the training data (Table 1), thus skewing predictions. TS (1wk) has the smallest error due to its ability to capture *changing and recent trends*. The RIM (Fig. 2c) with respect to DoW is greater than 0 for all models except TS (4wk), which performs worse than the baseline 7% of the time, even as it has comparatively smaller errors. Given the high *consumption variability* for RES, performance under volatility is important. A high VAB is desirable and provided by all models.

For OFF (Fig. 1d), we again see a divergence in model ranking when based on CVRMSE or on MAPE, reflecting the benefit of each measure. Uniquely, neither RT nor TS are able to surpass the DoW baseline in terms of CVRMSE. We independently verified if the consumption pattern of this building is highly-correlated with the DoW by examining the decision tree generated by RT, the best choice in terms of MAPE. We found the DoW feature to be present in the *root node of the tree* while the holiday flag was at the second level. RT is also the only model which (marginally) outperforms the baseline on RIM and VAB (Fig. 2d), thus delivering the benefits of using a feature-based approach that subsumes DoW. TS fails to do well, possibly due to *temporal dependencies* that extend beyond the 7-day lag period. It is notable that while DoW is the preferred model based on CVRMSE (Fig. 1d) for OFF, measures we propose, such as RIM and VAB (Fig. 2d) that evaluate performance against the DoW baseline, indicate that RT is the better choice.

For ACD (Fig. 1e), TS (1wk) and RT perform incrementally better than DoW on CVRMSE and MAPE, and VAB is positive for only these two models (Fig. 2e). The sharp change in standard deviation between the training and test data accounts for the higher sensitivity to volatility of the baseline (Table 1). But we observe slightly negative values of RIM for all models, implying *more frequent errors than the baseline*.

8.3 15-min Campus Predictions

The 15-min predictions for the campus shows TS (2hr) to fall closest to the observed values, based on CVRMSE (6.88%) and MAPE (4.18%) (Fig. 3a). This accuracy is validated relative to the baseline, with high RIM and VAB values (Fig. 4a). These reflect the twin benefits of *large spatial granularity* of the campus, which make its consumption slower changing, and the short horizon of TS (2hr), helping it capture *temporal similarity*. RT is the next best, performing similar to TS (6hr) and ToW baseline on CVRMSE, MAPE and RIM, though it is better with volatility (VAB=5.21%).

8.4 15-min Building Predictions

For 15-min predictions for buildings, we see that TS (2hr) is the only candidate model that always does better than the ToW baseline on all four measures (Figs. 3b–3e & 4b–4e). TS (6hr) and RT are occasionally better than ToW on CVRMSE and MAPE, and TS (24hr) rarely. Their CVRMSE errors are also uniformly larger than MAPE, showing that the models suffer more from occasional large errors. The academic environment with weekly class schedules encourages a uniform energy use behavior based on ToW, that is hard to beat. RES is the exception, where all candidate models are better than the baseline (Fig. 3c), given the aberrant summer months when it is unused.

However, when we consider the RIM and VAB measures, it is interesting to note that the candidate models are not significantly worse than the baseline (Fig. 4b–4e). In fact, TS (6hr) is better than ToW for all buildings but DPT, showing that it is more often accurate and more reliable under volatility. RT, however, is more susceptible to volatility and shows negative values for all buildings but RES. While TS (2hr) followed by TS (6hr) are obvious choices for short horizons, ToW and RT have the advantages of being able to predict over a longer term. In the latter case, ToW actually turns out to be a better model.

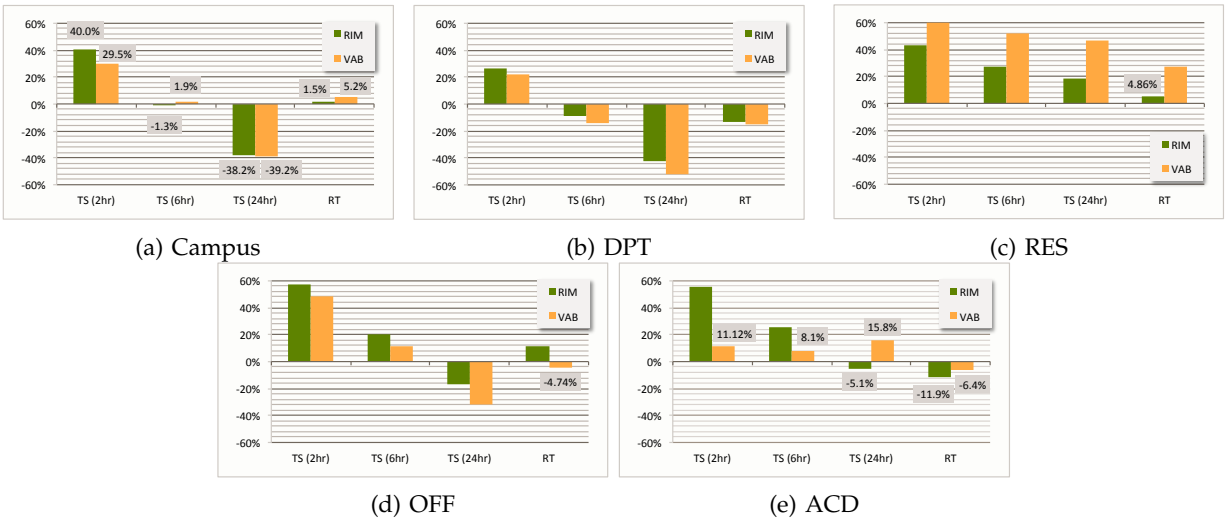
TABLE 2: **Application-independent Cost Measures**. Prediction horizon is 4 weeks for 24-hour predictions; and 24 hours for 15-min predictions. CD measures number of unique feature values used in training and testing. TS and baseline do not have a training cost.

Model	Data Cost CD	Compute Cost *(millisec)* CC_t	CC_p
DoW/ToW Baseline			
24-hour predictions	1,096	-	-
15-min predictions	1,05,216	-	-
Time Series			
24-hour predictions	1,096	-	101
15-min predictions	1,05,216	-	933
Regression Tree			
24-hour predictions	3,301	94	1.6
15-min predictions	1,31,629	17,275	48

8.5 Cost Measures

The data and compute cost measures, discussed here, are orthogonal to the other application-independent measures, and their values are summarized in Table 2. Making cost assessment helps grid managers ensure rational use of resources, including skilled manpower and compute resources, in the prediction process.

Data Cost (CD): The baselines and TS are univariate models that require only the electricity consumption values for the training and test periods. Hence their data costs are smaller, and correspond to the number of intervals trained and tested over. RT model has a higher cost due to the addition of several features (§ 7.2). However, the cost does not increase linearly with the number of features and instead depends on the number of unique feature values. As a result, its data cost is only $\sim 25\%$ and $\sim 300\%$ greater than TS for 24-hr and 15-min predictions respectively.



(a) Campus (b) DPT (c) RES

(d) OFF (e) ACD

Fig. 4: ***RIM*** and ***VAB*** values for ***15-min predictions*** for campus and four buildings. Higher values indicate better model performance with respect to the baseline; zero indicates similar performance as baseline. TS (2hr) usually offers highest reliability in all cases.

Compute Cost (CC): We train over 2 years and predict for 4 weeks (24-hour granularity) and 24 hours (15-min) on a Windows Server with AMD 3.0GHz CPU and 64GB RAM, and report the average over 10 experiment runs. The baseline's compute cost is trivial as it is just an average over past values, and we ignore it. For the TS models, retraining is interleaved with the prediction and we report them as part of the prediction cost (CC_p). We found prediction times for TS to be identical across campus and the four buildings, and the 15-min predictions to be $\sim 9\times$ the cost of 24-hour – understandable since there are $\sim 10\times$ the data points. The horizons did not affect these times. For RT, we find the training and prediction times to be similar (but not same) across campus and four buildings, and this is seen in the differences in the sizes of the trees constructed. We report their average time. While RT has a noticeable training time (17 secs for 15-min), its prediction time is an order of magnitude smaller than TS. As a result, its regular use for prediction is cheaper. It is more responsive, with a lower prediction latency, even as the number of buildings (or customers) increase to the thousands.

9 Analysis of Dependent Measures

The application-dependent measures (§ 5) enable model selection for specific application scenarios. For each application (§ 7.3), the measures' parameter values are defined in consultation with the domain experts. These values are listed in Tables 3 & 4.

9.1 Planning

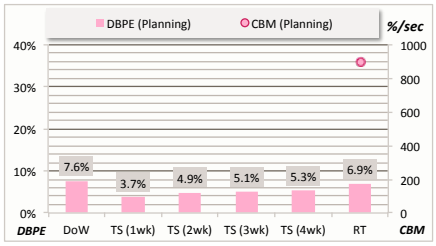
Planning requires medium- and long-term consumption predictions at 24-hour granularities for the campus and buildings, six times a year. The short horizon of TS (4 weeks) precludes its use. So we only consider DoW and RT models, but do report TS results.

TABLE 3: **Application Specific Parameters.** α, β are over- and under-prediction penalties for DBPE, and e_t is the error tolerance for REL.

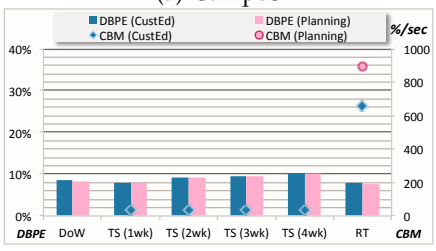
Application & Prediction Type	DBPE (α, β)	REL (e_t)
Planning		
24-hour Buildings	0.50, 1.50	0.15
24-hour Campus	1.00, 1.00	0.10
Customer Education		
24-hour Building	0.75, 1.25	0.15
15-min Buildings (6AM-10PM)	1.50, 0.50	0.10
Demand Response		
15-min Campus (1PM-5PM)	0.50, 1.50	0.05
15-min Buildings (1PM-5PM)	0.50, 1.50	0.10

Campus: For campus-scale decisions, both over- and under- predictions can be punitive. The former will lead to over-provisioning of capacity with high costs while the latter can cause reduced usability of capital investments. Hence, for DBPE, we equally weight $\alpha = 1$ and $\beta = 1$, whereby DBPE reduces to MAPE. We set $e_t = 10\%$, a relatively lower tolerance, since even a small swing in error % for a large consumer like USC translates to large shifts in kWh.

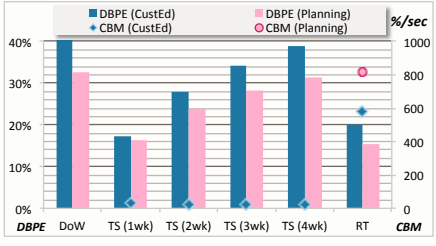
Fig. 5a shows RT (and TS) to perform better than the DoW baseline on DBPE (6.87% vs. 7.56%, consistent with MAPE). The RT model's reliability is also higher than DoW's (Fig. 6a), with RT providing errors smaller than the threshold 60.87% of the time – a probabilistic measure for the planners to use. When we consider the total compute cost for training and running the model (Table 4), RT is trained once a year and used six times, with a negligible compute cost of 103 msec and a high CBM of $900\%/sec$ (Fig. 5a). These make RT a better qualitative and cost-effective model for long term campus planning.



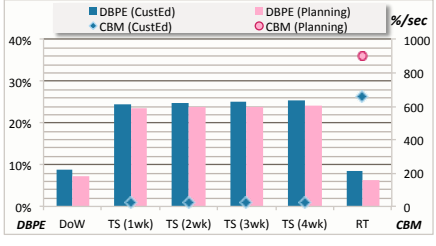
(a) Campus



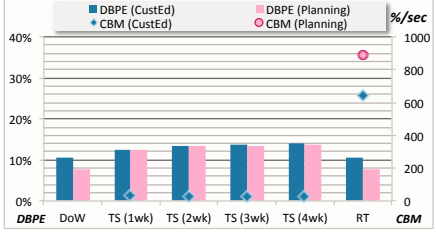
(b) DPT



(c) RES

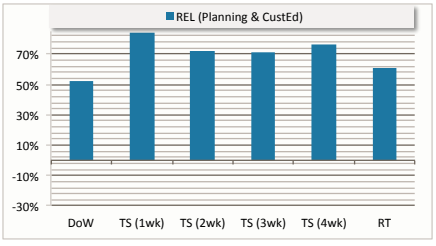


(d) OFF

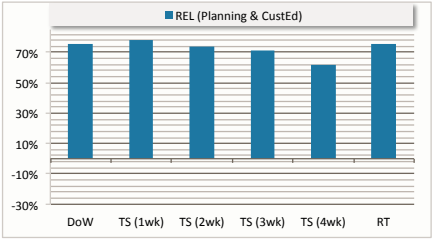


(e) ACD

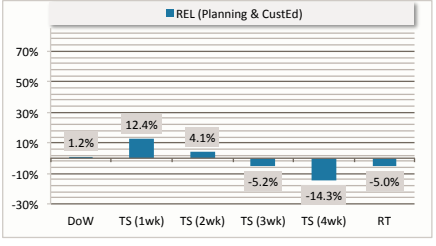
Fig. 5: Domain bias percentage error (***DBPE***), primary Y-axis, and Cost-Benefit Measure (***CBM***), secondary Y-axis, for ***24-hour predictions*** for Planning and Customer Education. Customer Ed. is not relevant for campus. Lower DBPE & higher CBM are better, as seen in RT.



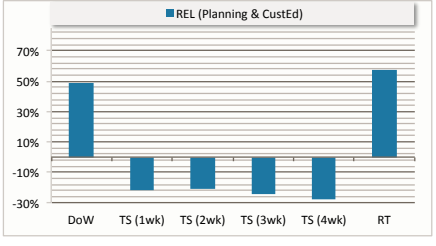
(a) Campus



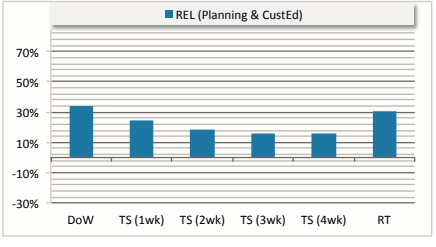
(b) DPT



(c) RES

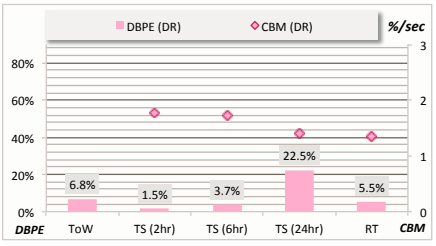


(d) OFF

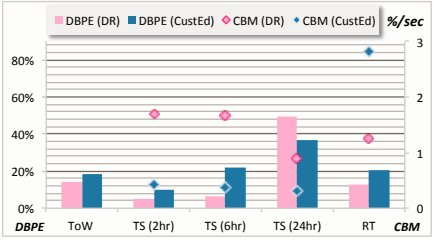


(e) ACD

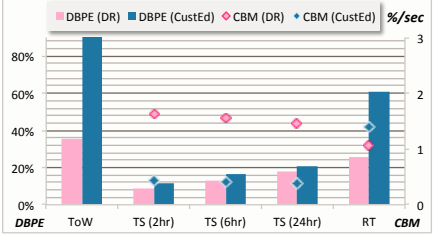
Fig. 6: Reliability (***REL***) values for ***24-hour predictions*** for Planning, and Customer Education. Both have the same value of error tolerance parameter, and shown by a single graph. Higher values indicate better performance than the baseline; zero matches the baseline.



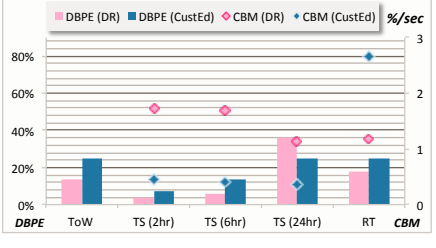
(a) Campus



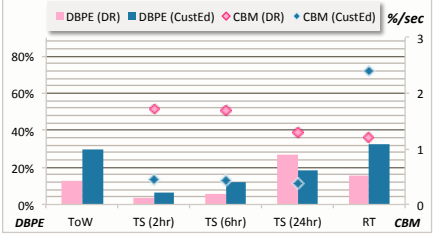
(b) DPT



(c) RES



(d) OFF



(e) ACD

Fig. 7: ***DBPE***, primary Y-axis, and ***CBM***, secondary Y-axis, for ***15-min predictions*** for Demand Response and Customer Education. Customer Ed. is not relevant for campus. Lower DBPE and higher CBM are desirable, and provided by TS (2hr) and TS (6hr) for DR.

Buildings: Buildings being upgraded for sustainability and energy efficiency favor over-prediction of consumption to ensure an aggressive reduction of carbon footprint. Reflecting this, we set $\alpha = 0.5$ and $\beta = 1.5$ for DBPE. A higher error tolerance than campus is acceptable, at $e_t = 15\%$. Cost parameters and measure values are the same as campus.

DBPE reflects a truer measure of error for the application and we see that it is smaller than MAPE across all models and buildings (Fig 5b-5e). Investigating the data reveals that the *average kWh for the training period was higher than that for the test period*, leading to over-predictions. Here, the models' inclination to over-predict works in their favor. While RT is uniformly better than DoW on DBPE, it is less reliable for RES and ACD (Figs. 6c & 6e), even falling below 0%, indicating that predictions go over the error threshold more often than below the threshold.

REL unlike DBPE treats over- and under-predictions similarly. While the baseline has fewer errors above the threshold, their magnitudes are much higher, causing DBPE (an average) to rise for smaller REL. The costs for RT are minimal like for campus and their CBMs similar. So the model of choice depends on if the predictions need to be below the threshold more often (DoW) or if the biased-errors are lower (RT). Particularly, for OFF, REL (Fig. 6d) shows RT is best for Planning even as DoW was the better model based on CVRMSE (Fig. 1d). Similarly, for ACD, REL (Fig. 6e) recommends DoW for Planning even as CVRMSE (Fig. 1e) suggests RT and MAPE (Fig. 1e) suggests TS (1wk). *These highlight the value of defining application-specific performance measures like REL for meaningful model selection.*

9.2 Customer Education

This application uses 24-hour and 15-min predictions at the building-level made during the daytime (6AM-10PM), and provides them to residents/occupants for monthly budgeting and daily energy conservation.

24-hour predictions: 24-hour predictions impact monthly power budgets, and over-predictions are better to avoid slippage. We pick $\alpha = 0.75$ and $\beta = 1.25$ for DBPE and an error tolerance $e_t = 15\%$ for REL. We use a 4-week prediction duration for costing with one 24-hour prediction done each day by RT and TS. RT is trained once in this period. We report TCC (Table 4), DBPE & CBM (Figs. 5b-5e), and REL (Figs. 6b-6e).

Like for Planning that preferred over-predictions, the DBPE here is smaller than MAPE for all models, and it is mostly smaller for RT and TS models than DoW. *But for a building like ACD, while one may have picked TS (1wk) based on the application-independent MAPE measure (Fig. 1e), both RT and DoW are better for Customer Education on DBPE (Fig. 5e).* Similarly, for both DPT and RES, TS (1wk) was the best option based on MAPE (Figs. 1b, 1c) as well as on DBPE for Customer Education (Figs. 5b, 5c). However, for a different application, such as Planning, RT is the recommended model based on DBPE (Figs. 5b, 5c). This highlights how a measure that is tailored for a specific application by setting tunable parameters can guide the effective choosing of models for it.

TABLE 4: **Application Specific Cost Parameters & TCC**. τ is the trainings per duration, and π is the model usage with a prediction horizon per duration.

Application	Trainings, τ	Uses, π (horizon)	TCC (millisec)
Planning (*duration = 1 year*)			
24-hour RT	1	6 (2mo)	103
Customer Education (*duration = 4 weeks*)			
24-hour RT	1	28 (1dy)	139
24-hour TS	-	28 (1dy)	2,845
15-min RT	1	8·28 (2hr)	28,103
15-min TS	-	8·28 (2hr)	2,09,037
Demand Response (*duration = 4 weeks*)			
15-min RT	4	5·3 (6hr)	69,824
15-min TS	-	4·5·3 (6hr)	55,992

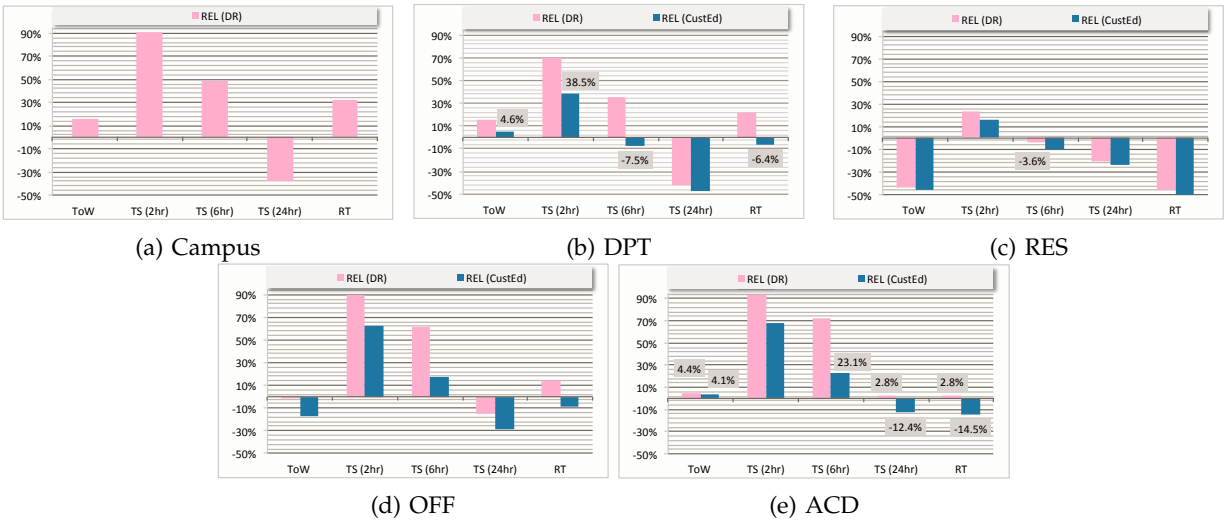
When considering reliability, REL for RT is marginally (DPT) or significantly (OFF) better than DoW even as the application-independent RIM showed RT to be worse or as bad as DoW respectively – yet another benefit of measures customized for the application. RT also equals or out-performs TS (1wk) on both DBPE and REL on all buildings but RES. The TCC cost for TS while being $\sim 20\times$ more than RT is still small given the one month duration. This also reflects in the CBM being much lower for TS.

15-min predictions: This application engages customers by giving periodic forecasts during the day to encourage efficiency. Over-predicting often or more frequent errors will mitigate a customer's interest. So we set $\alpha = 1.5$ and $\beta = 0.5$ for DBPE, and we have a lower error tolerance at $e_t = 10\%$ for REL. Prediction duration is 4-weeks for cost parameters, with 8 uses per day at 2 hour horizons. RT is trained once.

For all buildings, both DBPE and REL rank TS (2hr) as the best model (Figs. 7b-7e) & (Figs. 8b-8e). These reaffirm the effectiveness of TS for short-term predictions. For many models, the (daytime) DBPE for this application is higher than the (all-day) MAPE due to higher variations in the day. However, TS (2hr) bucks this trend for RES, OFF and ACD. RT is worse than even ToW on reliability, with REL below 0% for all buildings. For RES, all models but TS (2hr) have REL below 0%. So qualitatively, TS (2hr) is by far a better model. However, on costs (Table 4), TS has TCC ≈ 209 secs. This may not seem much but when used for 10,000's of buildings in a utility, it can be punitive. At large scales, CBM (Figs. 7b-7e) may offer a better trade-off and suggest RT for DPT and OFF.

9.3 Demand Response

DR uses 15-min predictions to detect peak usage and preemptively correct them to prevent grid instability. Hence, over-predictions are favored than under to



(a) Campus (b) DPT (c) RES

(d) OFF (e) ACD

Fig. 8: ***REL*** values for ***15-min predictions*** for Demand Response and Customer Education. Higher is better.

avoid missing peaks, and we set $\alpha = 0.5$ and $\beta = 1.5$ for DBPE. The campus is a large customer with tighter requirements of error threshold at $e_t = 5\%$ for REL, while individual buildings with lower impact are allowed a wider error margin of $e_t = 10\%$. Prediction duration is 4 weeks for cost parameters, with the models used thrice a weekday – before, at the start and during the 1-5PM period, and RT trained weekly.

DBPE is uniformly smaller than MAPE for the campus and buildings (Figs. 7a-7e), sometimes even halving the errors. Thus the 4 hour DR periods in the weekdays are more (over-)predictable than all-day predictions. TS (2hr) has significantly better DBPE than other models, with even TS (6hr) out-performing RT and ToW. For campus, RT is better than DoW, in part due to using temperature features that have a cumulative impact on energy use during midday.

We see TS (2hr) gives a high REL of 91% for campus (Fig. 8a) and is the only model with positive REL for RES. Also, TS (6hr) and RT prove to be more reliable for DR in campus and DPT than their poorer showing in the RIM and VAB independent measures (Fig. 4), making them competitive candidates. However, RT suffers in reliable predictions for other buildings, with lower or negative REL (Figs. 8c-8e) while TS (2hr) and (6hr) continue to perform reliably.

Cost-wise, we see RT and TS models are comparable on TCC (Table 4). For once, RT takes longer than TS due to the more aggressive retraining (every week), preferred for critical DR operations. But when seen through the CBM measure, all TS models beat RT for all cases but one (TS (24hr) on DPT). Thus, the TS (2hr) and TS (6hr) are the best for DR on all measures.

10 Conclusion

The key consideration in evaluating a prediction model by an end-user is its performance for the task at hand. Traditionally, accuracy measures have been used as the sole measure of prediction quality. In this article, we examine the value of holistic performance measures along the dimensions of scale independence, reliability and cost. In evaluating them for consumption prediction in Smart Grids, we see that *scale independence* ensures that performance can be compared across models and applications and for different customers; *reliability* evaluates a model's consistency of performance with respect to baseline models; while *cost* is a key consideration when deploying models at large scale for real world applications. We use existing scale-independent measures, *CVRMSE* and *MAPE*, while extending and proposing four additional measures, *RIM* and *VAB* for measuring reliability; and *CD* and *CC* for data and compute costs.

Further, our novel application-dependent measures can be customized by domain experts for meaningful model evaluation for applications of interest. These measures include *DBPE* for scale independence, *REL* for reliability, and *TCC* and *CBM* for cost. The value of these measures for scenario-specific model selection were empirically demonstrated using three Smart Grid applications that anchored our analysis even as they are generalizable to other domains. Through cross correlation analysis, we found that only MAPE and CVRMSE show absolute correlation > 0.9, indicating that all measures are individually useful. Our results demonstrate the valuable insights that can be gleaned on models' behavior using holistic measures. These help to improve their performance, and provide an understanding of the predictions' real impact in a comprehensive yet accessible manner. As such, they offer a common frame of reference for model evaluation by future researchers and practitioners.

Acknowledgments

This material is based upon work supported by the United States Department of Energy under Award

Number DEOE0000192, and the Los Angeles Department of Water and Power (LA DWP). The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, the LA DWP, nor any of their employees.

REFERENCES

[1] Emergency Demand Response Program Manual, Sec 5.2: Calculation of Customer Baseline Load (CBL). Technical report, New York Independent System Operator (NYISO), 2010.

[2] H. K. Alfares and M. Nazeeruddin. Electric load forecasting: literature survey and classification of methods. *International Journal of Systems Science*, 33(1):23–34, 2002.

[3] S. Aman, Y. Simmhan, and V. K. Prasanna. Improving energy use forecast for campus micro-grids using indirect indicators. In *IEEE Workshop on Domain Driven Data Mining*, 2011.

[4] N. Amjady. Short-term hourly load forecasting using time-series modeling with peak load estimation capability. *IEEE Transactions on Power Systems*, 2001.

[5] J. Armstrong and F. Collopy. Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, 8(1):69–80, 1992.

[6] G. E. P. Box and G. M. Jenkins. *Time series analysis, forecasting and control*. Holden-Day, 1970.

[7] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman and Hall, 1984.

[8] P. Chakraborty, M. Marwah, M. Arlitt, and N. Ramakrishnan. Fine-grained photovoltaic output prediction using a bayesian ensemble. In *AAAI Conference on Artificial Intelligence*, 2012.

[9] B.-J. Chen. Load forecasting using support vector machines: a study on EUNITE competition 2001. *IEEE Transactions on Power Systems*, 2004.

[10] R. Cottet and M. Smith. Bayesian modeling and forecasting of intraday electricity load. *Journal of the American Statistical Association*, 98(464):839–849, 2003.

[11] K. Coughlin, M. A. Piette, C. Goldman, and S. Kiliccote. Estimating demand response load impacts: Evaluation of baseline load models for non-residential buildings in California. Technical Report LBNL-63728, Lawrence Berkeley National Lab, 2008.

[12] A. Davydenko and R. Fildes. Measuring forecasting accuracy: The case of judgmental adjustments to sku-level demand forecasts. *International Journal of Forecasting*, 29(3), 2013.

[13] B. Dong, C. Cao, and S. E. Lee. Applying support vector machines to predict building energy consumption in tropical region. *Energy and Buildings*, 37:545–553, 2005.

[14] C. W. J. Granger. Prediction with a generalized cost of error function. *Operational Research Quarterly*, 20:199–207, 1969.

[15] J. D. Hobby and G. H. Tucci. Analysis of the residential, commercial and industrial electricity consumption. In *IEEE Innovative Smart Grid Technologies Asia (ISGT)*, 2011.

[16] R. Hyndman and A. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 2006.

[17] R. J. Hyndman and Y. Khandakar. Automatic time series for forecasting: The forecast package for R. Technical report, Monash University, 2007.

[18] IBM Software Information. Managing big data for smart grids and smart meters. Technical report, IBM Corporation, 2012.

[19] N. Jewell, M. Turner, J. Naber, and M. McIntyre. Analysis of forecasting algorithms for minimization of electric demand costs for EV charging in commercial and industrial environments. In *Transportation Electrification Conference and Expo*, 2012.

[20] G. Kenny, A. Meyler, and T. Quinn. Forecasting Irish inflation using ARIMA models. Technical Report 3/RT/98, Central Bank of Ireland, 1998.

[21] A. Khotanzad, R. Afkhami-Rohani, and D. Maratukulam. ANNSTLF-Artificial Neural Network short-term load forecaster generation three. *IEEE Transactions on Power Systems*, 13(4), 1998.

[22] O. Kramer, B. Satzger, and J. Lassig. Power prediction in smart grids with evolutionary local kernel regression. *Hybrid Artificial Intelligence Systems, LNCS*, 6076:262–269, 2010.

[23] F. Martinez-Alvarez, A. Troncoso, J. Riquelme, and J. A. Ruiz. Energy time series forecasting based on pattern sequence similarity. *IEEE Transactions on Knowledge and Data Engineering*, 23(8), 2011.

[24] J. L. Mathieu, D. S. Callaway, and S. Kiliccote. Variability in automated responses of commercial buildings and industrial facilities to dynamic electricity prices. *Energy and Buildings*, 2011.

[25] A. McKane, I. Rhyne, A. Lekov, L. Thompson, and M. Piette. Automated demand response: The missing link in the electricity value chain. *ACEEE Summer Study on Energy Efficiency in Buildings*, 2008.

[26] H. Mori and A. Takahashi. Hybrid intelligent method of relevant vector machine and regression tree for probabilistic load forecasting. In *IEEE International Conference and Exhibition on Innovative Smart Grid Technologies (ISGT Europe)*, 2011.

[27] R. C. Prati, G. E. A. P. A. Batista, and M. C. Monard. A survey on graphical methods for classification predictive performance evaluation. *IEEE Transactions on Knowledge and Data Engineering*, 23(11), 2011.

[28] S. D. Ramchurn, P. Vytelingum, A. Rogers, and N. R. Jennings. Putting the 'smarts' into the smart grid: A grand challenge for artificial intelligence. *Communications of the ACM*, 55(4):86–97, 2012.

[29] J. Shishido. Smart meter data quality insights. In *ACEEE Summer Study on Energy Efficiency in Buildings*, 2012.

[30] Y. Simmhan, S. Aman, A. Kumbhare, R. Liu, S. Stevens, Q. Zhou, and V. Prasanna. Cloud-based software platform for data-driven smart grid management. *IEEE/AIP Computing in Science and Engineering*, 2013.

[31] Y. Simmhan, V. Prasanna, S. Aman, S. Natarajan, W. Yin, and Q. Zhou. Towards data-driven demand-response optimization in a campus microgrid. In *ACM Workshop On Embedded Sensing Systems For Energy-Efficiency In Buildings (BuildSys)*, 2011.

[32] M. Sokolova, N. Japkowicz, and S. Szpakowicz. Beyond Accuracy, F-Score and ROC: A family of discriminant measures for performance evaluation. In *Australian Joint Conference on Artificial Intelligence*, 2006.

[33] L. F. Sugianto and X. Lu. Demand forecasting in the deregulated market: a bibliography survey. *Australasian Universities Power Engineering Conference*, pages 1–6, 2002.

[34] L. Torgo and R. Ribeiro. Precision and recall for regression. In *International Conference on Discovery Science*, 2009.

[35] T. Verbraken, W. Verbeke, and B. Baesens. A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Transactions on Knowledge and Data Engineering*, 25(5), 2013.

[36] K. Wagstaff. Machine learning that matters. In *International Conference on Machine Learning (ICML)*, pages 529–536, 2012.

[37] X. Yu, C. Cecati, T. Dillon, and M. G. Simes. The new frontier of smart grids. *IEEE Industrial Electronics Magazine*, 2011.

Saima Aman is a PhD student in Computer Science at the Univ of Southern California and a Research Assistant in the Smart Grid project at the Center for Energy Informatics. Her research interests are in the areas of Data Mining and Artificial Intelligence. She has a Master's in Computer Science from the University of Ottawa.

Yogesh Simmhan is an Assistant Professor in the Indian Institute of Science, and previously a faculty at the University of Southern California and a postdoc at Microsoft Research. His research is on scalable programming models for Big Data applications on distributed systems like Clouds. He has a PhD in Computer Science from Indiana University. Senior member of IEEE and ACM.

Viktor K. Prasanna is the Powell Chair in Engineering, Professor of Electrical Engineering and Computer Science, and Director of the Center for Energy Informatics at the Univ of Southern California. His research interests include HPC, Reconfigurable Computing, and Embedded Systems. He received his MS from the School of Automation, Indian Institute of Science and PhD in Computer Science from Penn State. He is a Fellow of IEEE, ACM and AAAS.