Summary of Research on
Accuracy of Peak Load Predictions
September 10, 2020

Jessica Granderson, Eliot Crowe, Samir Touzani

Lawrence Berkeley National Laboratory

Overview

- # Research Questions

  - How accurately do weather regression methods and matching methods predict peak loads in commercial buildings?

  - What are the implications for demand response and load shed programs?

Methodology

Build dataset of hourly load and weather data for meters with no known energy efficiency improvements or DR events

For each meter define model training period (baseline) and prediction period (hypothetical event window)(10x per meter)

Use models to predict hourly load during hypothetical event windows and compare to actual load; calculate error metrics

Repeat for all meters in dataset, for each prediction method of interest; quantify distribution of error metrics for each method

Compare predictive accuracy of commonly-used DR impact estimation methods with weather regression methods

Test Dataset

- 858 meters, hourly data on electricity, temperature
- Prediction/event days selected from most recent 365 days for each meter
  - Chose the 10 days with highest maximum load
  - Two hypothetical event windows: 10am–6pm and 12pm-6pm
  - Weekends and holidays excluded
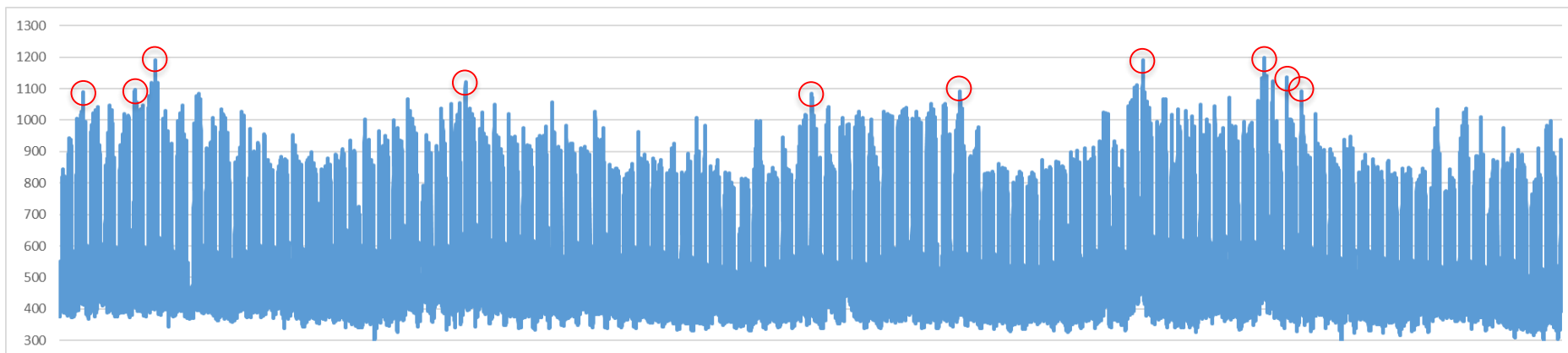
Commonly Used DR Estimation Methods

- North American Energy Standards Board (NAESB)
  - **Maximum Base Load:** uses system load and individual meter data from past DR season to generate flat, constant level of demand for baseline that customer must remain at or below.
  - **Meter Before / Meter After:** uses actual load data from time period immediately preceding event
  - **Baseline Type-I:** uses historical interval meter data and may also use weather data to generate baseline.
  - **Baseline Type-II:** uses statistical sampling to generate baseline for portfolio of customers in instances where interval meter for all individual sites is not available
  - **Metering Generator Output:** baseline set as zero and measured against usage readings from behind-the-meter emergency back-up generators. Only applicable for facilities with on-site generators

Most commonly-used in current DR programs

BERKELEY LAB
Lawrence Berkeley National Laboratory

Baseline Type-I

- Variations include Averaging ("High X of Y"), Regression, Rolling Average, Comparable Day

- A High X of Y baseline considers 'Y' most recent days preceding an event and uses data from 'X' of these Y days to calculate baseline

- 2 variants:
  - **Day-Matching:** Subset of non-event days in close proximity to event day are identified/averaged
  - **Weather Matching:** Similar to above except baseline load profile based on non-event days with similar temperature
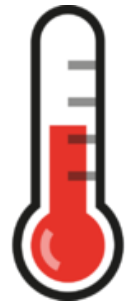
Matching Methods Assessed

- ## Day-Matching:
  - Average 10 of last 10 eligible days (weekdays immediately prior to event)

- ## Weather Matching:
  - Average 4 weekdays from 90 days immediately prior to event, with closest maximum daily temperature to event day

Recommendations from California ISO Baseline Assessment, 2017 (Nexant)

Weather Regression Methods Assessed

- Time of Week and Temperature (TOWT), weighted (recent days more influence) and unweighted

- Dynamic Regression with ARIMA Errors:
  - Regression based on Temperature with ARIMA errors
  - Regression based on TOWT with ARIMA errors

BERKELEY LAB
Lawrence Berkeley National Laboratory

Baseline Modeling

- **Baselining methods:**
  - Day-matching (10 days)
  - Weather-matching (4 days from 90)
  - TOWT:
    - Baseline days: 7; No weighting
    - Baseline days: 70; Weighting: 14 days
    - Baseline days: 70; Weighting: 10 days
  - Adjustments for 'pre-event' load (option for all methods except weighted TOWT)

- **Number of model runs***
  - 12:00-6:00 event window: 4,907
  - 10:00-6:00 event window: 4,840

* Maximum theoretical 8,580 model runs (10 events for each of 858 buildings), but some permutations could not meet minimum data requirements to run models

BERKELEY LAB
Lawrence Berkeley National Laboratory

Illustration of Method (TOWT with 7-day baseline)

This day had one of the top ten highest peak demands for this meter

10:00-6:00 on this day is selected as a hypothetical "event window"

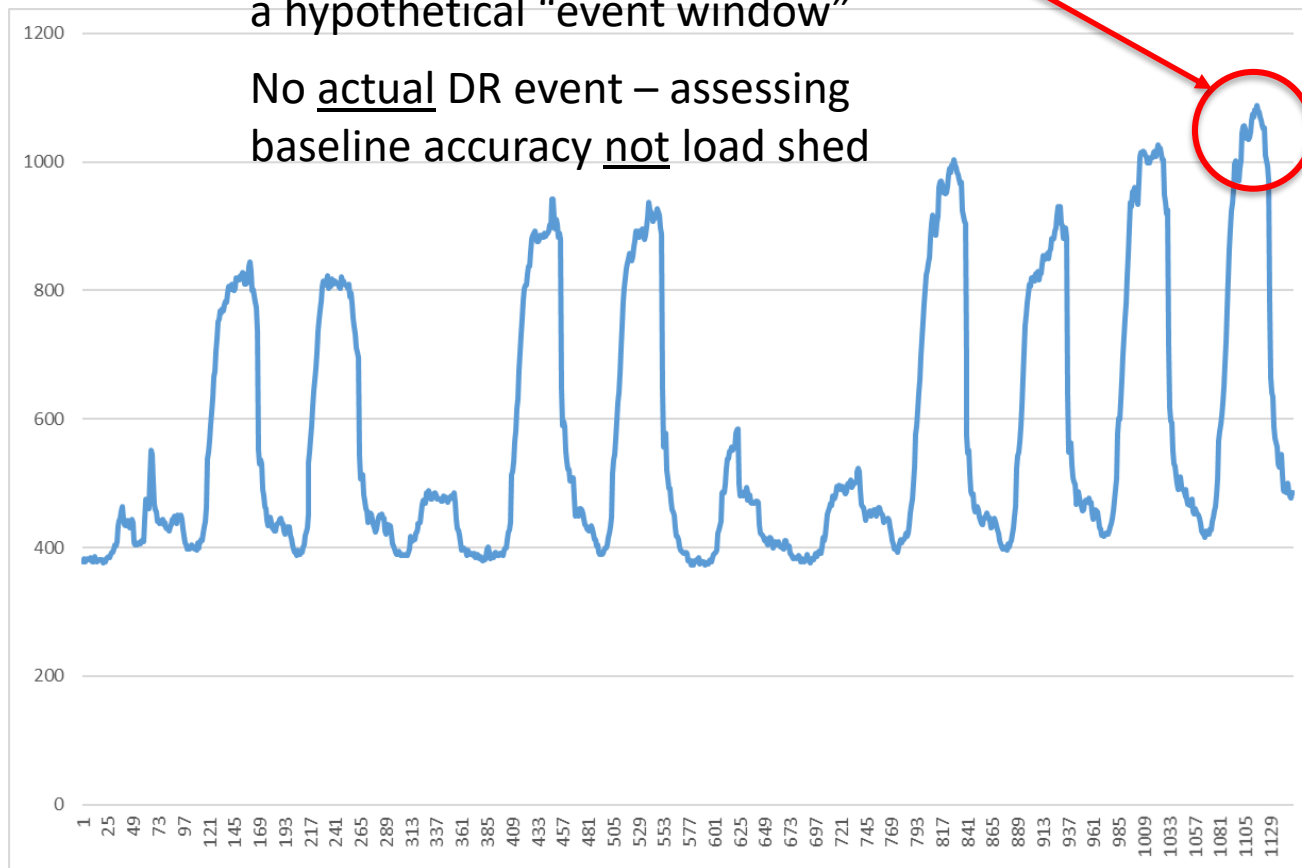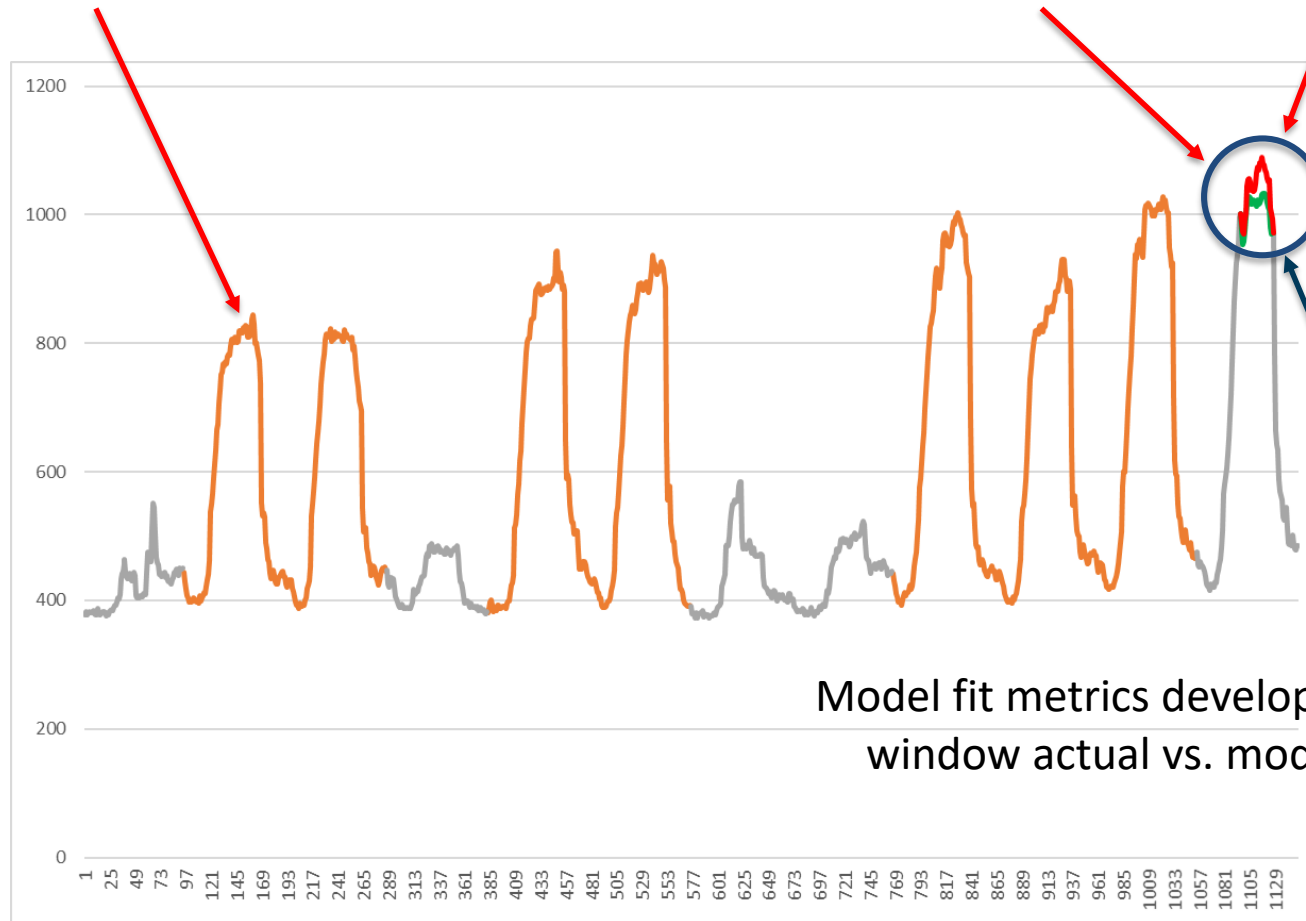No <u>actual</u> DR event – assessing baseline accuracy <u>not</u> load shed

Illustration of Method (TOWT with 7-day baseline)

7 prior workdays used to train TOWT model (orange)

TOWT predicts hourly load for event window (green)

Actual event window load (red)

Model fit metrics developed using event window actual vs. model-predicted

BERKELEY LAB
Lawrence Berkeley National Laboratory

Assessment Metrics

Normalized mean bias error, $NMBE = \dfrac{\frac{1}{N}\sum_i^N(y_i - \hat{y}_i)}{\bar{y}} \times 100$

CV of the root mean squared error, $CV(RMSE) = \dfrac{\sqrt{\frac{1}{N}\sum_i^N(y_i - \hat{y}_i)^2}}{\bar{y}} \times 100$

- Provide a complement in understanding model performance

- NMBE is total percent difference between predicted and actual energy use (zero is best)

- CV(RMSE) indicates model's ability to predict the overall load shape (lower is better)
  - familiar to practitioners
  - prominent in resources such as ASHRAE Guideline 14

BERKELEY LAB
Lawrence Berkeley National Laboratory

Results: Accuracy Metric: NMBE

Day matching & Weighted
TOWT: Lower median values,
and tighter distribution
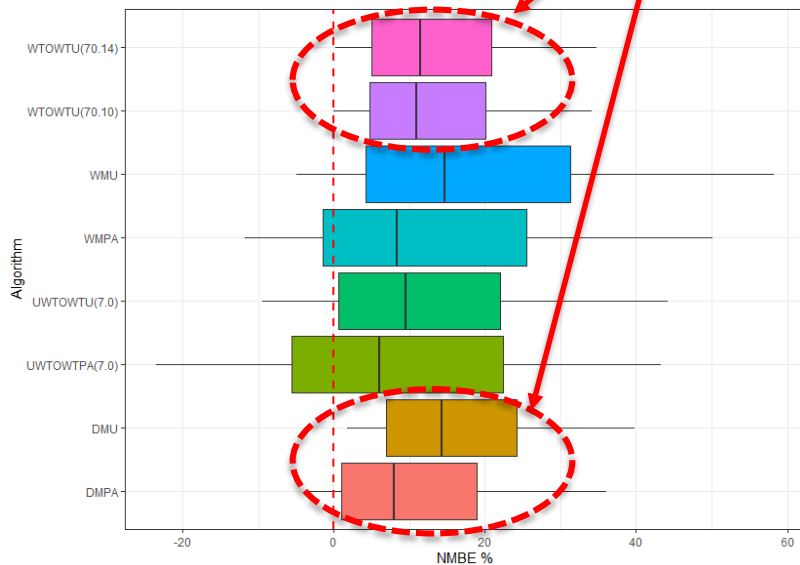
10:00-6:00 Event window

- NMBE of zero is best
- All methods tend to underpredict through 'event' window
- Pre-event adjustment reduced bias, improved prediction
- Variations in median and distributions, but high degree of overlap between methods
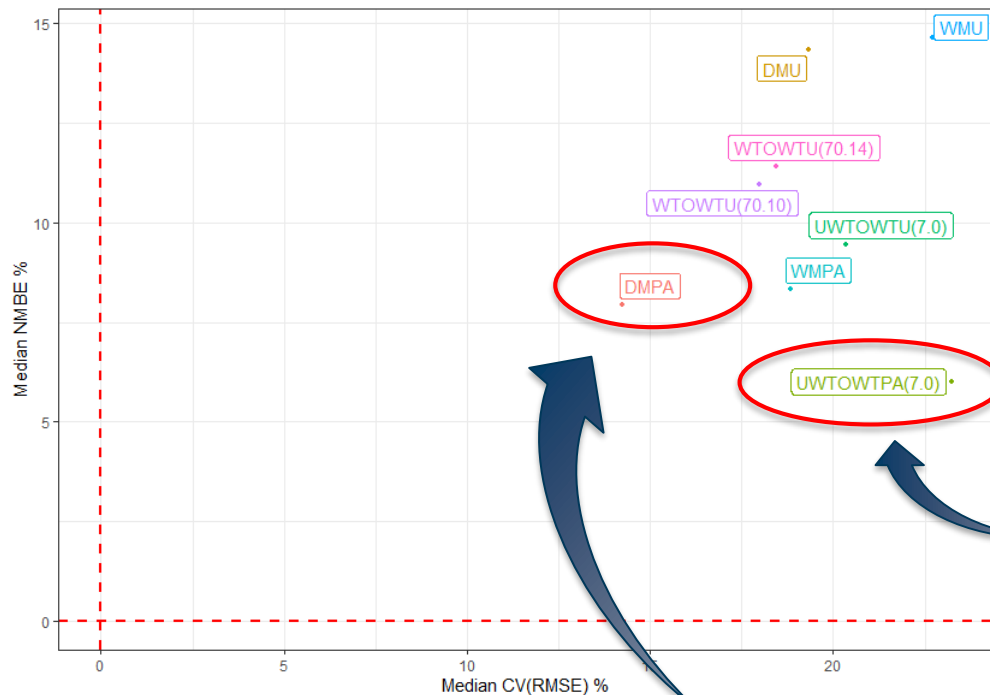
Results: CV(RMSE)

Day matching & Weighted
TOWT: Lower median values,
and tighter distribution
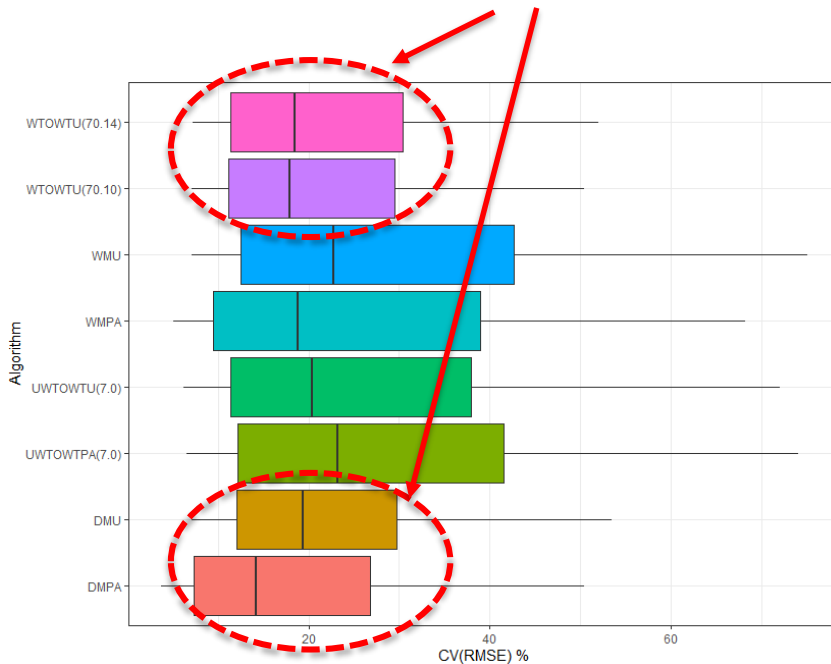
10:00-6:00 Event window

- Lower values are better

- As with NMBE, high degree of overlap

- Weather-matching and unweighted TOWT have very wide distribution
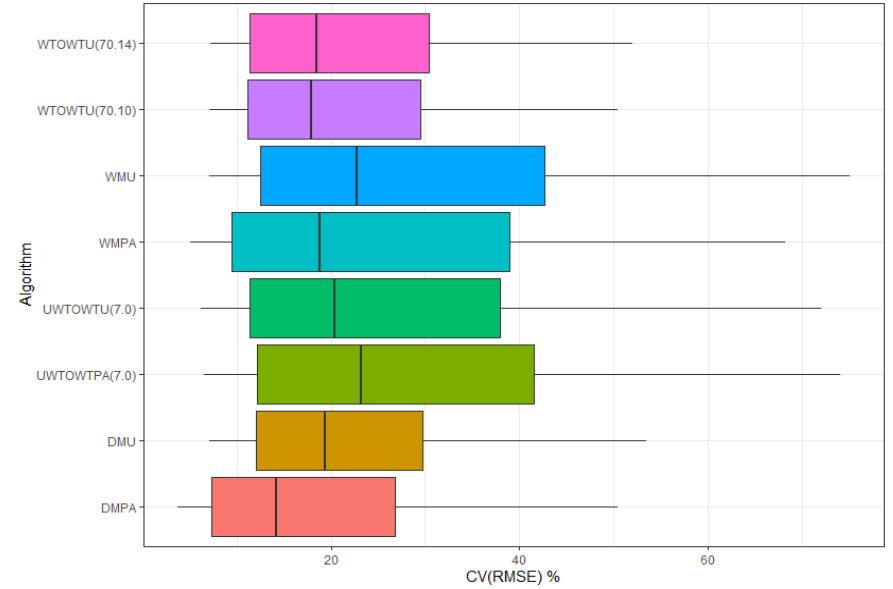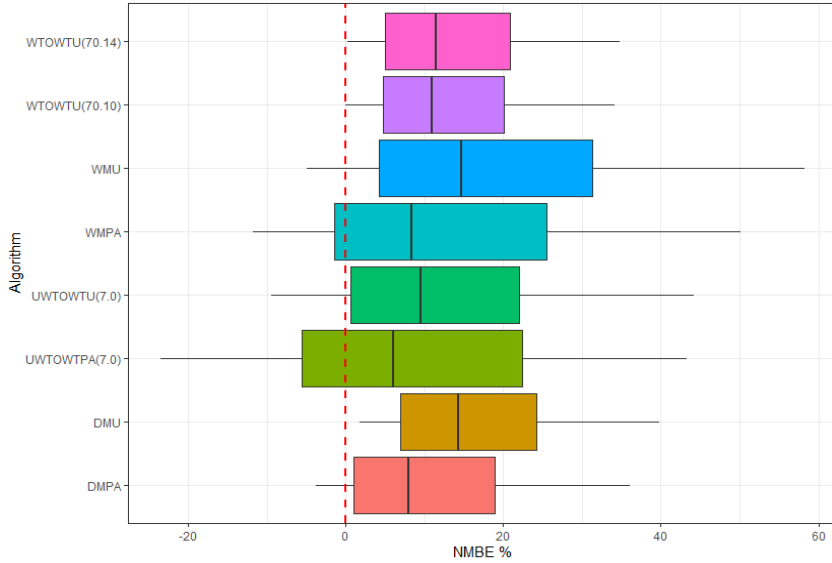
Median NMBE vs. CV(RMSE)

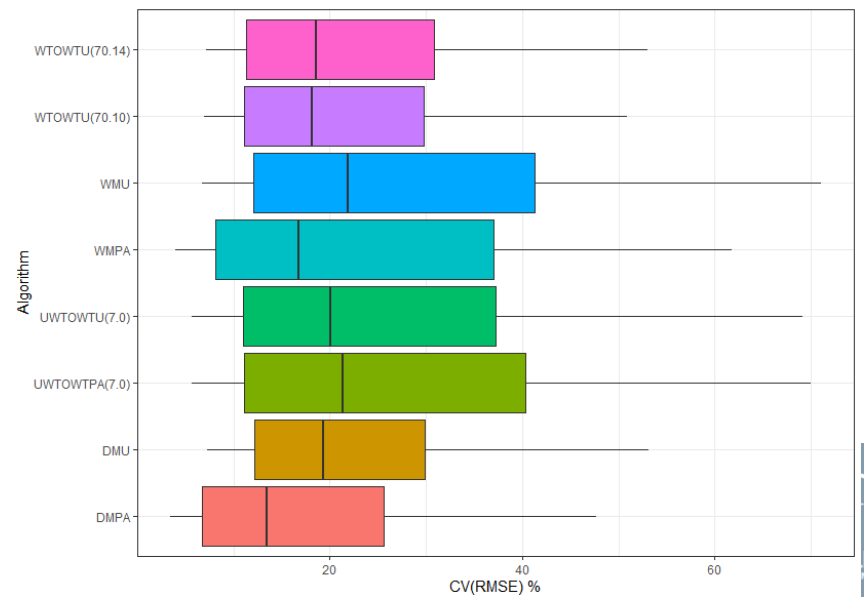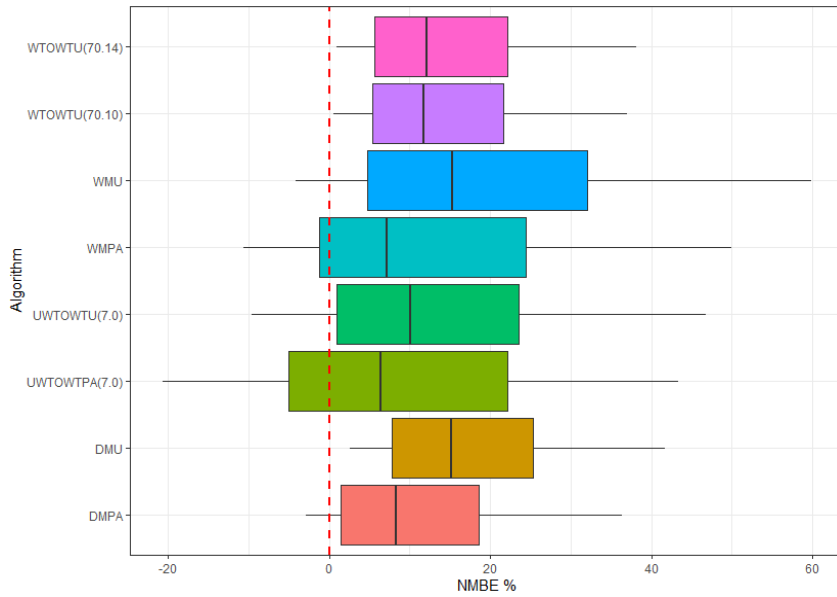- Unweighted TOWT with 7 day baseline and adjustment was least biased (NMBE)

- Day-Matched adjusted combined lowest CV(RMSE) and near-lowest bias).

- WMPA and WTOWT next-best if considering both metrics

BERKELEY LAB
Lawrence Berkeley National Laboratory

Comparing Event Windows

Conclusions

- All methods underpredicted load, i.e., will underestimate peak load sheds (median NMBE values ranged from 6.0% – 15.2%)

- CV(RMSE) median values ranged from 13.5% - 23.2%

- Absolute differences in medians are small, differences in performance surface in *distributions* of error

- Regression-based methods offered no clear advantage over averaging methods

- Weighted TOWT performed better vs. unweighted

- Adjustments based on pre-event conditions reduced bias

- 'Best' method depends on how ones values:
  - NMBE vs. CV(RMSE) as evaluation metrics
  - Lower median value vs. tighter distribution

BERKELEY LAB
Lawrence Berkeley National Laboratory

Improving Prediction of Peak/High Loads

- Consider other modeling methods that have shown promise in other domains (FY21)

- Consider other independent variables that might improve peak prediction
  - Rarely are these available at scale from measurements, but might test in simulation

BERKELEY LAB
Lawrence Berkeley National Laboratory

Discussion

- Heteroscedasticity is acknowledged, but results still surprising
- Implications of results?
  - Individual settlements
  - Aggregators/program-level/evaluation
- Relative emphasis on NMBE vs. CV(RMSE)?
- Other relevant factors?
  - Influence of peak load vs peak temperature
  - Alignment of system peak and building peak
  - Selection of event time windows

BERKELEY LAB
Lawrence Berkeley National Laboratory