# Protocol Development for Demand Response Calculation

# Draft Findings and Recommendations

**Prepared for**

**California Energy Commission**
**Sacramento, California**

**WORK AUTHORIZATION #1**
Contract 400-00-062

**Prepared by**

**XENERGY Inc.**
**Madison, Wisconsin**

**August 1, 2002**

**ma:project:wcec0007 dr support:report:5baseline recommendations:draft to CEC REVISED:title**

XENERGY Inc. • 2001 W. Beltline Highway • Suite 200 • Madison, WI 53713 • (608) 277-9696

# X EXECUTIVE SUMMARY

## X.1 INTRODUCTION

### X.1.1 Background on this Project

During the electricity crises of the last two years, a number of states and utilities within these states have developed programs to encourage customers to reduce their peak loads on short notice (under 2 to 24 hours) in exchange for some form of compensation. Such demand response (DR) programs depend on a credible operational procedure for determining the magnitude of load reductions.

The use of inconsistent methods for calculating baselines and corresponding load reductions has caused both confusion and dissatisfaction among participating customers. The lack of a standard measurement procedure may be reducing the number of customers willing to participate in DR programs, particularly in smaller- and medium-sized commercial customers in California.

### X.1.2 Objectives

The objective of this work is to develop a standardized measurement and verification (M&V) protocol for use by building engineers, facility operators, or outside M&V experts to "measure" the load drops achieved at a premises. Completion of this protocol is aimed at increasing participation in DR programs from small- and medium-sized customers by reducing the barriers related to inconsistency and confusion about baseline methods.

### X.1.3 Project Steps

Steps in the project include:

- Review of existing methods
- Testing of alternative methods on data sets from various locations and customer types
- Draft report on findings and recommendations, circulated for review and presented for discussion at a public workshop
- Final report
- Submission of the final recommendations to the International Performance Measurement and Verification Protocol (IPMVP) organization for adoption as part of the IPMVP.

This is the draft report on findings and recommendations. Included are the review of existing methods and the results of tests on alternative methods.

### X.1.4 The Role of the IPMVP

The IPMVP organization has participated in the development of this report and recommendations. The organization is responsible for the continued development and dissemination of standardized verification methods. It is hoped that the involvement of the IPMVP at various stages of review and the anticipated adoption of the DR protocol as an IPMVP document will represent a broad base of support for the framework developed.

There are direct parallels in the current demand response area to what was occurring in the world of M&V for energy-efficiency performance contracting eight years ago. The core concept of the IPMVP document is that parties involved in contracts to reduce energy use should have a common language with which to structure and manage the settlement of those contracts. The IPMVP was designed to allow parties flexibility in designing M&V procedures that make sense for each contact.

As is true for the energy-efficiency IPMVP, the intent of this report is not to provide a prescriptive set of steps and rules. Rather, the goal is to establish a clear vocabulary, and to offer guidelines on good practice and the pros and cons of alternative method specifications. Toward the goal of developing consistent terminology, this document develops a taxonomy of different methods and attempts to provide clear definitions. We anticipate that a discussion of definitions and distinctions will be an important part of the refinement of this document.

### X.1.5 The Role of Other Contributors

This work would not have been possible without the contributions of several other organizations and individuals.

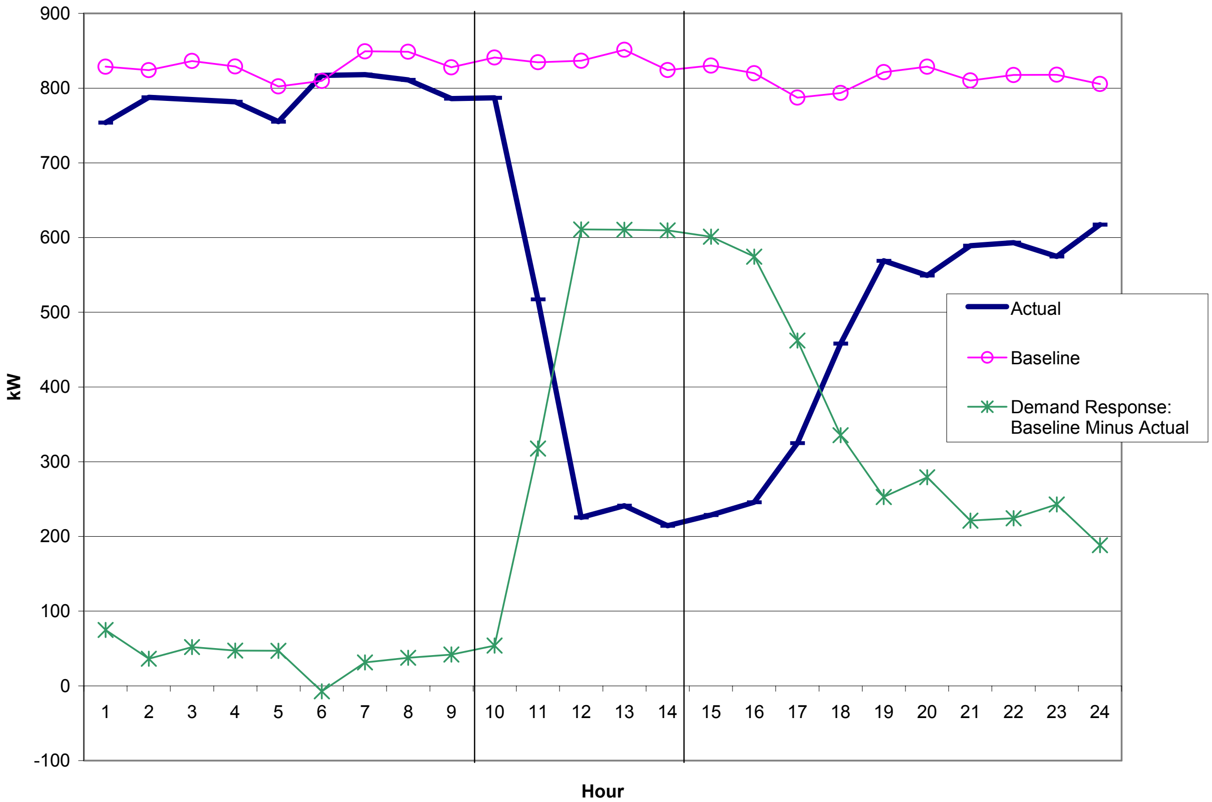
- *Method donors* have shared details of methods they have developed and applied for quantifying demand reductions.
- *Data donors* have provided interval load data from curtailed and noncurtailed customers for use in testing methods.
- *Sponsors* have provided some funding in addition to supplying data and/or methods.
- *Reviewers* are reviewing the major project deliverables.

## X.2 REVIEW OF EXISTING METHODS

### X.2.1 The Demand Response Baseline

The focus of this study was on calculations of demand response from whole-premise interval load data. Demand response is calculated as the difference between the *baseline* and the actual metered load in each interval (Figure X-1). The baseline is the estimate of what the load would have been in each interval in the absence of the curtailment. Thus, the key question for the demand response calculation is how the baseline is determined.

**Figure X-1**
**Example of Demand Response Calculation from Baseline and Actual Loads**




### X.2.2 Desirable Features in a Demand Response Baseline Calculation Method

Most of the goals developers described for the baseline were similar. They were

- to reflect load that would have been used absent the program,
- ease of use for program participants,
- ease of use for program administrators, and
- deterrence of gaming.

Given the multiple objectives, all customer baseline developers understood that the baseline methodology they chose was a compromise. Criteria that are balanced in developing a baseline include

- Simplicity
- Ease of use
- Ease of understanding

- Verifiability
- Accuracy
- Lack of bias (i.e., no systematic tendency to over- or under-state reductions)
- Ability to handle weather-sensitive accounts fairly
- Minimization of gaming
- Ability to be known prior to customer's commitment to a particular curtailment amount and event
- Costs for participant and operator to implement
- Consistency with other ISOs.

### *X.2.3 Components of Whole-Premise Baseline Calculation Methods*

Baseline calculation methods based on whole-premise interval-metering data can be described in terms of three fundamental components:

- **Data selection criteria** determine what days and time periods of data will be used in the baseline calculation.
- The **estimation method** is a calculation procedure that determines the provisional baseline load at each interval for the curtailment day, using the data selected by the data selection criteria.
- The **adjustment method** shifts or scales the provisional baseline to align it with known conditions of the curtailment day.

#### *Data Selection Criteria*

Common starting points for data selection include

- Use of the last 10 to 20 uncurtailed business days
- Use of a subset of the last 10 or 11 business days that had the highest load
- Use of a full season of data.

Selection criteria include varying procedures for excluding days from the starting point and replacing excluded days, sometimes in an iterative process.

#### *Estimation Method*

Most estimation methods can be characterized as either an average or some form of weather-based regression model.

### *Adjustment Methods*

Common adjustment methods include:

- Unadjusted
- Additive
- Scalar
- Weather-based additive or scalar.

## *X.2.4 Characterization of Existing Methods*

Existing baseline estimation methods are summarized in Figure X-1 in terms of the components discussed above.

## Figure X-2
## Summary of Existing Baseline Estimation Methods

| | Load Type Differences Addressed* | How Weather Sensitivity is Addressed** | Time Interval | Data Selection: Initial Timeframe | Data Selection: Final Selection | Data Selection: Excluded Days (other than weekends and previous program Control days) | Estimation Method | Adjustment Hours, If Adjusted |
|---|---|---|---|---|---|---|---|---|
| PJM-Day Ahead 2002 | WS/NWS, Self Gen, Cust. Spec. | Top 5 of 10, Optional adjustment to control-day weather | Hourly | 10 days, n-2 to n-11 | High 5 of 10 | Low Output Days. | Interval Average | h-1,h-2. |
| PJM Emergency | None | None | Hourly | Hour before | Same | None | | |
| ISO-NE 2001-2002 | WS/NWS, Self Gen, | Adjustment to control-day load | Hourly | 10 days, n-1 to n-10 | Same | Extreme Output Days. | Interval Average | h-1,h-2. |
| NYISO-DADRP 2001 | Self Gen | Top 5 of 10 | Hourly | 10 days, n-2 to n-11 | High 5 of 10 | Low Output Days. | Interval Average | |
| NYISO-DADRP 2002 | WS/NWS, Self Gen | Top 5 of 10, Optional adjustment to control-day load | Hourly | 10 days, n-2 to n-11 | High 5 of 10 | Low Output Days. | Interval Average | h-3, h-4. |
| ERCOT-BUL 2002 | WS/NWS | Optional adjustment to control-day load | 15 minute | 10 days, n-1 to n-10 | Middle 8 | None | Interval Average | h-1, h-2. |
| CAISO 2001#1 | None | None | Hourly | 10 days | None | None | Interval Average | |
| CAISO 2001#2 | None | None | Hourly | 11 days | None | None | Interval Average | |
| XENERGY | WS/NWS | Regression-based estimate, Adjustment to control-day load | Hourly | Variable | None | None | Regression-based | h-1, h-2. |
| LBNL/Kinney | WS/NWS | Regression-based estimate | Hourly | 10 days, n-1 to n-10 | None | None | Regression-based | |
| Nexant | WS/NWS | Adjustment to control-day load | 15 minute | 10 days | None | None | Interval Average | h-1 |
| Utility A | WS/NWS | Adjustment to control-day load | Hourly | Previous Month | None | All Days that do not fit the match-day criteria. | Interval Average | one hour, 8am - 11am |
| Utility B | None | None | Hourly | 5 days | None | Customer-specificed anomalous loads | Interval Average | |
| Utility C | WS/NWS | Regression-based estimate | Hourly | Undefined minimum data | None | None | Regression-based | |
| Utility D | WS/NWS | Regression-based estimate | Hourly | Weekdays, June through September | None | None | Regression-based | 5am - 10am |
| Utility E | None | Match based data selection | 15 minute | Undefined | 10 Days with min. SSE compared to day n-1 | None | Interval Average | All match-day hours. |
| Utility F | WS/NWS, Cust. Spec. | Adjustment to control-day load | Hourly | 2-3 previous years | None | Anomalous loads | Interval Average | h-1, h-2. |
| CMTA Proposed OBMC | WS/NWS | Adjustment to control-day load | Hourly | 10 days, n-1 to n-10 | None | None | Interval Average | h-1 through h-4 |

* WS/NWS: Different methods for weather-sensitive and nonweather-sensitive loads
Self Gen: Different methods for onsite generation

** Top 5 of 10: Select 5 days with highest average load during the hours curtailed on the curtailment day

## X.3 FINDINGS FROM METHOD TESTS

Several combinations of data selection criteria, estimation method, and adjustment were tested on interval load data from curtailed and uncurtailed accounts across the country. A total of 646 accounts were included in the tests. For accounts that were not curtailed, baseline estimates were compared with actual load for each hour of an actual or simulated curtailment period. For accounts that were curtailed, each candidate baseline estimate was compared with the estimate produced by the "best" method.

Performance was assessed in terms of both bias and overall error magnitude. Bias is the systematic tendency to over- or under-state the baseline and corresponding demand reduction. Variability is how wide the swings are around the typical or expected value. Overall error magnitude reflects both bias and variability. Key findings are indicated below.

### *X.3.1 Adjustments*

- Additive adjustment to the two hours before curtailment can reduce the bias and variability of almost all methods, including weather models, for all load types. Other types of adjustments can improve the performance of averages, but generally with higher bias and variability.
- With this additive adjustment, simple averages can perform essentially as well as complex weather models, even for weather-sensitive accounts.
- Without adjustment, most averages tend to understate load.
- Additive adjustment to the last 2 hours can be problematic for several reasons:
    - It opens the possibility of gaming by deliberately increasing load just before the curtailment period to boost the baseline.
    - Legitimate pre-cooling in response to a curtailment notice or expectation will also erroneously increase the baseline.
    - Conversely, an operation that achieves its curtailment target promptly upon notification and before the beginning of the required curtailment period will have a severely understated baseline.

### *X.3.2 Data Selection*

- Bias and variability of weather models is reduced by longer input data series, but not dramatically.
- The decreased variability with longer input series is more noticeable for conditional weather models applied to non-weather-sensitive accounts, particularly high-variability loads.

- The different averages compared performed similarly in terms of bias and variability, except for those that select a subset of days based on high load.
- For summer loads, the High 5 of 10 average reduces the otherwise negative bias. For summer loads using additive adjustment, High 5 of 10 gives the lowest bias of any of the averages, for both weather-sensitive and non-weather-sensitive accounts, and comparable variability. The High 10 of 11 average gives some bias reduction, but not as much.
- For nonsummer loads, however, the High 5 of 10 average inflates an already positive bias. The other averages perform better and roughly comparably to each other, in terms of both bias and variability, for both weather-sensitive and non-weather-sensitive accounts. The High 10 of 11 is somewhat better than the others.

### *X.3.3 Weather Modeling*

- For summer weather-sensitive accounts, weather models tend to perform somewhat better than averages, but the difference is not dramatic.
- For summer non-weather-sensitive accounts, use of a "conditional" weather model does not increase bias or variability. The conditional weather model automatically deletes weather terms if the load data indicate they are inappropriate for a particular account. Thus, if weather models are used, a single methodology can be applied to both weather-sensitive and non-weather-sensitive accounts.
- For nonsummer loads, weather models do not perform better than averages.

## X.4 PROS AND CONS OF ALTERNATIVE APPROACHES

Advantages and disadvantages of key method features in terms of the criteria indicated in Section X.2.2 are summarized in the table below. This table is based on both qualitative considerations and the results of the performance tests.

# Table X-1
## Advantages and Disadvantages of Key Baseline Method Features Based on Qualitative Considerations and Test Results

| Baseline Method | Variant | Pros | Cons |
|---|---|---|---|
| Average | Any | Simple, easy to use and understand, low cost | Tends to understate baseline for weather-sensitive loads, especially if unadjusted |
| | High 5 of last 10 days | Partial adjustment for weather-sensitive loads | Still tends to understate baseline for weather-sensitive loads |
| | | | Can allow windfall load reduction credit on cool days |
| Regression | Any | Provides baseline corresponding to particular weather conditions of curtailment day | More complex, harder to understand, higher cost |
| | | | If observations don't include conditions as extreme as the curtailment day, model estimate may be inaccurate |
| | | | If account isn't weather-sensitive, may be less accurate than simpler methods |
| | Full Season | Adequate data and range of variation to yield accurate coefficients | Operating conditions from the period data are taken from may be different from curtailment day |
| | Recent 10 days | Operating conditions more likely to be similar to curtailment day | Model based on limited data may be inaccurate |
| | Lag temperature/ degree-day | Tends to reduce bias for weather-sensitive accounts | Tends to increase variability of baseline estimate. |
| | Conditional | Allows same general form and procedure to be used for weather-sensitive and non-weather-sensitive accounts, without pre-screening. Doesn't add much error for non-weather-sensitive accounts. | More complex. May give less consistent results across events for an account, if weather terms are sometimes retained and sometimes not. |
| Adjustment to precurtailment hours | Any | Simple, easy to use and understand, low cost | May be potential for gaming behavior during day-of-curtailment adjustment period |
| | | Adjusts to weather and operating conditions of curtailment day | Appropriate pre-curtailment increase in load (e.g., pre-cooling) will result in overstated baseline |
| | | Limits potential for collecting windfall credits for planned shut-downs | Pre-curtailment decrease in load in response to curtailment request (e.g., long ramp-down, canceling a shift) will result in understated baseline |
| | Additive | May adjust well for load change that is constant throughout day (e.g., industrial processes) | May not be appropriate if load changes during curtailment period (ratio adjustment may be better suited) |
| | Scalar | May adjust well for load change that is function of exogenous factor throughout day (e.g., higher levels of occupancy) | May not be appropriate if the day-to-day load variation is constant over the day (additive adjustment may be better suited) |
| | to last 2 hours before curtailment period | If load in these hours is unaffected by anticipated or initiated curtailment, provides best accuracy | If substantial curtailment is initiated in these hours, severely understates baselines |
| | to 3rd and 4th hour before curtailment period | Less potential for understated baseline due to pre-curtailment-period demand response | More variability than adjustment to last 2 hours |
| Weather-Based Adjustment | Any | Explicitly takes into account weather conditions | Adjustment may not be known to customer until after curtailment period (i.e., until after weather conditions are known for the day) |
| | | No opportunity for gaming as with adjustment to precurtailment hours | If no observations are available for extreme conditions, estimates used for adjustment may be outside range of model |
| | | | Will badly predict load reductions if the buildings are dominated by internal loads |
| | | | Less accurate than alternative adjustments or weather model for both weather-sensitive and non-weather-sensitive accounts |

## X.5 RECOMMENDATIONS

### *X.5.1 Proposed Approaches by Load Type*

#### *Offering Options*

A general recommendation is that baseline calculation protocols should provide for alternatives based on customer load types and operating practices. One way to simplify the provision of options is to establish a default method and allow certain deviations.

The basis for the selection of a method should be not just the business type, but also the load patterns evident in the data as well as the customer's description of operating practices. Thus, for example, a customer who indicates a desire to be able to cancel a shift in advance of the control period should have access to a baseline calculation method that is not distorted by this practice.

At the same time, the program operator should have some discretion to bar customers from using an approach that they appear to have manipulated in the past. Thus, if there is evidence that a particular customer tends to inflate load after notification, beyond what would reasonably be expected for pre-cooling, that customer might not be able to use a method that includes adjustment to the 2 pre-curtailment hours.

#### *A Practical Default Baseline Calculation Method*

A method that generally works well for a range of load types is the simple average of the last 10 days, with additive adjustment to the 2 hours prior to the curtailment period. This method can be recommended for both weather-sensitive and non-weather-sensitive accounts, with both low and high variability, for summer and nonsummer curtailments.

This method is not recommended for accounts that tend to curtail in advance of the required period in response to a curtailment notice. It is also not recommended for situations where the potential for gaming is a strong concern, whether across the program or for particular customers.

#### *Alternatives for Summer Weather-Sensitive Accounts*

For summer programs, practical alternatives for weather-sensitive accounts include the following:

- Unadjusted weather models. Longer input time periods are preferable, particularly for high-variability loads.
- The High 5 of 10 average with THI adjustment.

Simpler methods with less desirable but potentially acceptable performance include:

- Unadjusted averages, particularly the High 5 of 10.

- Averages or weather models adjusted to the third and fourth hour before curtailment.

### *Alternatives for Summer Non-Weather-Sensitive Accounts*

For non-weather-sensitive summer loads, the unadjusted High 10 of 11 average performs nearly as well as the recommended default, particularly for low-variability loads. Next best is the simple average of the last 10 days with additive adjustment to the third and fourth hours before curtailment.

For low-variability loads, unadjusted weather models, with weather terms retained only if indicated by the data, actually perform slightly better than the recommended default. However, unlike the case for weather-sensitive accounts, these models perform better if based on shorter periods of data. For high-variability loads, unadjusted weather models tend to be worse than the unadjusted High 10 of 11 average.

### *Alternatives for Nonsummer Accounts*

For nonsummer loads, modeling is more challenging and there are fewer alternatives. For weather-sensitive accounts, the High 5 of 10 average with THI adjustment can be used. For low-variability loads, the unadjusted High 5 of 10 appears to perform slightly better, but for high-variability loads it is worse.

For non-weather-sensitive nonsummer loads, the unadjusted High 10 of 11 appears to be the best alternative. Any of the averages with additive adjustment to the third and fourth hour before curtailment do not perform as well.

### *Summary of Recommended Methods and Alternatives*

The recommended methods and alternatives for different load types are summarized in the table below.

**Table X-2**
**Recommended Methods and Alternatives**

| Season | Weather Sensitivity | Variability | Recommended Default: Estimation | Recommended Default: Data Selection | Recommended Default: Adjustment | Recommended Alternatives: Estimation | Recommended Alternatives: Data Selection | Recommended Alternatives: Adjustment |
|---|---|---|---|---|---|---|---|---|
| Summer | Weather-Sensitive | Low | Average | last 10 | add 1-2 | weather models | any | none |
| | | | | | | Average | High 5 | THI |
| Summer | Weather-Sensitive | High | Average | last 10 | add 1-2 | weather models | longer is better | none |
| | | | | | | Average | High 5 of 10 | THI |
| Summer | Non-Weather-Sensitive | Low | Average | last 10 | add 1-2 | weather models | shorter is better | none |
| | | | | | | Average | High 10 of 11 | none |
| | | | | | | Average | last 10 | add 3-4 |
| Summer | Non-Weather-Sensitive | High | Average | last 10 | add 1-2 | Average | High 10 of 11 | none |
| | | | | | | Average | last 10 | add 3-4 |
| Nonsummer | Weather-Sensitive | Low | Average | last 10 | add 1-2 | Average | High 5 of 10 | none |
| Nonsummer | Weather-Sensitive | High | Average | last 10 | add 1-2 | Average | High 5 of 10 | THI |
| Nonsummer | Weather-Sensitive | Low | Average | last 10 | add 1-2 | Average | High 10 of 11 | none |
| | | | | | | Average | last 10 | add 3-4 |
| Nonsummer | Weather-Sensitive | High | Average | last 10 | add 1-2 | Average | High 10 of 11 | none |
| | | | | | | Average | last 10 | add 3-4 |

# 1 INTRODUCTION

## 1.1 PROJECT BACKGROUND

During the electricity crises of the last two years, a number of states and utilities within these states have developed programs to encourage customers to reduce their peak loads on short notice (under 2 to 24 hours) in exchange for some form of compensation. Compensation may be a specified incentive payment per kW reduced, or other benefits. Many programs involve a contracted magnitude demand reduction, with incentives paid if the contracted load drop is delivered. Some involve a penalty if the contracted reduction is not delivered.

Such demand response (DR) programs depend on a credible operational procedure for determining the magnitude of load reductions. Different regulatory jurisdictions, utilities, and system operators have ended up defining different methods of calculating both the load curtailments achieved by participating accounts and the “baseline” load shapes that should be used to calculate the level of load reduction realized.

The use of inconsistent methods for calculating baselines and corresponding load reductions has caused both confusion and dissatisfaction among participating customers. This situation may serve as a barrier to entry to new customers who want to participate in demand response programs but don’t want to take the time to master the details of estimating and confirming load reductions at their premises. More importantly, the lack of a standard measurement procedure may be reducing the number of customers willing to participate in DR programs, particularly in smaller and medium sized commercial customers in California.

The objective of this work is to develop a standardized measurement and verification (M&V) protocol for use by building engineers, facility operators or outside M&V experts to “measure” the load drops achieved at a premise. Completion of this protocol is aimed at increasing participation in DR programs from small and medium sized customers by reducing the barriers related to inconsistency and confusion about baseline methods.

The intent of the protocol is to facilitate program participation and operations. As a result, the protocol development must consider the practicality of implementation and potential effects on program and customer operations as well as the technical performance of candidate methods. To this end, this work begins with a review of the issues that drive the development of demand response baseline methods, and the rationale behind several existing methods. We then provide the technical performance assessment for a wide range of methods. The recommendations developed attempt to strike a balance between technical accuracy and practical implications for program operators and participants.

## 1.2 PROJECT AND REPORT ORGANIZATION

### *1.2.1 Steps in the Project*

Project tasks include:

- Review of existing methods
- Selection of alternative methods for testing
- Testing of alternative methods on data sets from various locations and customer types
- Draft report and workshop on findings and recommendations
- Compilation of comments on the draft
- Final report
- Submission of the final recommendations to the International Performance Measurement and Verification Protocol (IPMVP) organization for adoption as part of the IPMVP.

This is the draft report on findings and recommendations. Included are the review of existing methods and the results of tests on alternative methods.

### *1.2.2 Organization of the Report*

The next section provides background on demand response calculation methods. Included here are discussions of the purposes of the calculation, and the rationale for different method features. We provide a taxonomy of demand response calculation methods to provide a consistent vocabulary for discussing these features.

In Section 3 we discuss the issues and concerns that have been at the forefront in the development of demand response calculation methods in different parts of the country. We describe the specific approaches that have been adopted to address these concerns, and summarize advantages and disadvantages of key method features. Appendix A provides further detail on individual methods.

Section 4 describes the analysis conducted to test alternative method performance. Results of the analysis are presented in Section 5. Recommendations based on the review and analysis are offered in Section 6. A glossary of terms is given in Appendix C.

### *1.2.3 The Role of the IPMVP*

The International Performance Measurement and Verification Protocol (IPMVP) organization has participated in the development of this report and recommendations. The "IPMVP " refers to a document as well as an organization. The organization is responsible for the continued development of the concept of standardized verification methods, distribution of the document, and providing guidance and training on the appropriate use of the document. The latest document (Concepts and Options for Determining Energy and Water Savings) discusses issues pertaining to quantifying the long-term results of energy efficiency projects. The core concept of the IPVMP is that parties involved in contracts to reduce energy use should have a common language with which to structure and manage the settlement of those contracts. The IPMVP was designed to allow parties flexibility in designing monitoring and verification (M&V) procedures that make sense for each contact.

There are direct parallels in the current demand response area to what was occurring in the world of M&V for energy efficiency performance contracting eight years ago. Development of protocols for DR calculation methods with the goal that they may be adopted as another IPMVP document means several things. First, two individuals representing the IPMVP organization are actively involved in this project. They bring to the project the lessons and discipline of the earlier IPMVP development. Second, the approach to demand response measurement protocols is similar to that of the energy efficiency IPMVP. That is, the intent is not to provide a prescriptive set of steps and rules. Rather, the goal is to establish a clear vocabulary, and to offer guidelines on good practice and the pros and cons of alternative method specifications. Finally, the involvement of the IPMVP organization at various stages of review and the anticipated adoption of the DR protocol as an IPMVP document will represent a broad base of support for the framework developed.

Toward the goal of developing consistent terminology, this document develops a taxonomy of different methods, and attempts to provide clear definitions. A glossary of terms and acronyms is included as Appendix C. We anticipate that a discussion of definitions and distinctions will be an important part of the refinement of this document.

### *1.2.4 The Role of Other Contributors*

Several other organizations have contributed to this work in one of several forms:

- *Method donors* have shared details of methods they have developed and applied for quantifying demand reductions.
- *Data donors* have provided interval load data from curtailed and noncurtailed customers for use in testing methods.
- *Reviewers* have (agreed to) review the major project deliverables.

The contributors at the different levels are listed below. *Reviewers will be acknowledged in the final report.*

**Table 1-1**
**Contributors**

*†Brian Soth, Director, Retail Energy Products, and Meghan Jonee-Guinn, Operations Manager, Demand Buy Back Program Portland General Electric Company
*†Carl Raish, Administrator, Load Research, Tampa Electric
*Tennessee Valley Authority
*†Suzanne Galster, Load Research Manager, and Brett Johnson, Load Research Contractor/Analyst, Xcel Energy
Additional data and method donors, names withheld by request.

* Data Donor
† Method Donor

# 2 BACKGROUND

This section provides background on demand response calculation methods. We begin by discussing the different purposes that a demand response calculation may serve. We then present a "taxonomy" of demand response calculation methods. We describe broad classes of approaches to determining individual customers' demand reductions during particular control events. We then present a classification scheme for methods based on analysis of whole-premise load data. These methods are the focus of this study.

## 2.1 PURPOSES OF DEMAND RESPONSE CALCULATION METHODS

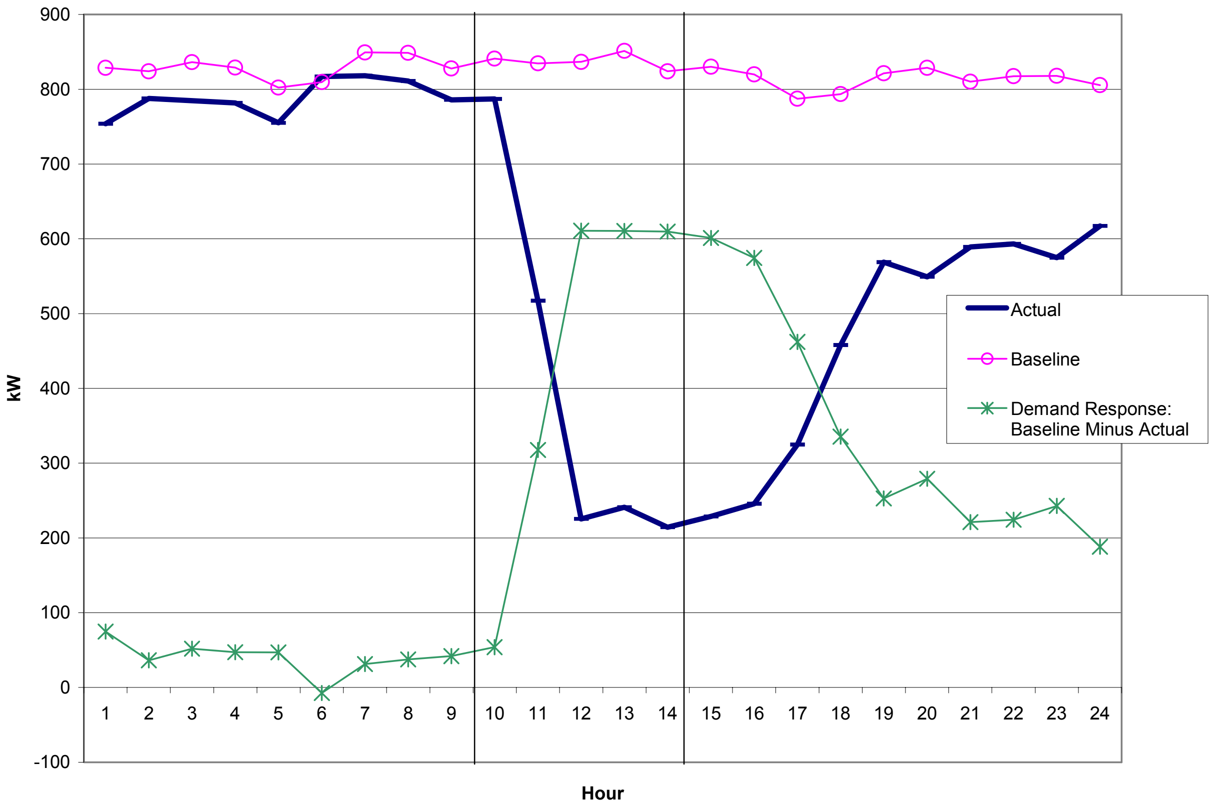
Programs that pay customers according to the amount of demand reduced during each curtailment period require an agreed procedure for calculating this demand response. The objective of this study is to assess alternative methods for this application.

Figure 2–1 illustrates the demand response calculation. The actual load is the customer's metered load for each hour of the day. This load shows a precipitous drop in response to a curtailment call. The "baseline" is an estimate of what the customer's load would have been across the day in the absence of the curtailment call. This baseline matches the actual load reasonably well prior to the curtailment. The difference between the baseline and the actual load is the calculated demand response for each hour of the curtailment period. Thus, the load reduction credited to the customer depends on the method for calculating the baseline estimate of what would have occurred in the absence of the program.

There are load management programs that do not require explicit calculation of each customer's load reduction as a basis for financial settlement between the customer and the utility or program operator. Utility programs based on firm load agreements or stipulated default loads, described below, are primary examples. For these programs, customer load reductions are often calculated for purposes of overall program evaluation and rate setting. Several of the methods reviewed here were developed for this purpose, rather than as a formal basis for financial settlement for program participation.

Demand response calculation methods differ also in the parameters targeted by the calculation. For programs that make payments based on calculated load reduction in each curtailment period, the key parameter is the difference between the actual load (kW) and load that would have occurred absent the program kW for each time increment in the curtailment period. In some cases, the minimum, maximum, or average of these differences over the curtailment period may be of interest. For evaluation, rate setting, and sometimes for determination of capacity requirements in capacity markets, the demand reduction at the time of the system peak is important. Evaluation and payments may also require estimates of total energy reduction (kWh) over control periods as well as demand reduction in each interval.

**Figure 2-1**
**Example of Demand Response Calculation from Baseline and Actual Loads**




Many of the newer demand response programs offered by Independent System Operators (ISOs) provide monetary incentives to end-users commensurate with actual load reductions. This is a shift from traditional utility load management programs that attract participation with lower tariffs in return for the potential of a certain number of service curtailments. Demand reduction for the ISO programs generally involves computation of the kW reduction for each time increment of the control period. In some cases (e.g., the New York ISO) there are separate protocols for measuring load (kW) demand response as compared to energy (kWh) response, as the demand reduction is bid into two separate markets. The development of a baseline protocol is then at least a partial function of its purpose.

Our review of existing methods and their development focuses primarily on methods developed for ISO or utility programs that attempt to measure the whole-premise demand reduction for each time increment of the control period. Deviations from this context are explicitly noted.

## 2.2 TAXONOMY OF DEMAND RESPONSE CALCULATION METHODS

### *2.2.1 Classes of Approach*

Crediting customers for demand response or load reduction during a given period requires a means of determining the amount of load reduced. There are three primary classes of approach to the demand response calculation.

A. **Stipulated Default Load**. The most common example of a stipulated approach is a "firm load" arrangement. A participant reduces load to a pre-determined "firm" load level upon notice. Failure to do so may invoke penalties, expose the participant to market prices, or requires the participant to pay for the energy at a higher than usual price. Determination of payments due to and from the customer depends only on confirmation that the customer's load did not exceed the firm level during the control period. The credit given to the customer for reducing to the firm level on notice is negotiated in advance, based on an estimate of the typical magnitude of load reduction.

   Another version of a stipulated default allows the customer to specify and pay for a base load level. The customer then pays for load above that level or receives credit for load below that level at prices related to market conditions. This approach provides a variant of a real-time pricing program.

   These approaches are analogous to the "stipulated" level A in the energy efficiency IPMVP. The payment due to the provider of energy savings is based on a pre-specified or stipulated savings magnitude.

B. **End-Use Metering of Self-Generation**. The demand response is measured as the metered onsite generation. There may be no attempt to account for what level of onsite generation might otherwise have occurred, or to assess whether load at the site has increased in ways that offset the additional generation.

   This approach is analogous to level B of the energy efficiency IPMVP, which uses end-use metering to assess savings for particular subsystems.

C. **Whole Premise**. The demand response is measured as the difference between the metered load at each interval and an estimate of the load that would have occurred absent the demand response program. This estimate of load in the absence of the program is referred to by most programs and in this document as the "baseline" or "customer baseline load" (CBL). The baseline is determined from analysis of whole-premise interval metering data before and after the beginning of the control period. The demand response can be a combination of demand reduction and/or increased use of onsite generation. In either case the net load is metered at the whole premise utility or meter service provider revenue meter.

This approach is analogous to level C of the energy efficiency IPMVP. Under that approach, whole-premise metered energy consumption from before and after the energy efficiency measures were implemented is analyzed to determine the reduction associated with the efficiency measure.

Use of stipulated approaches and metering of self-generation are fairly straightforward and have not been highly contentious. The focus of this review is on the Whole Premise approach, and in particular on baseline calculation methods.

### *2.2.2 Components of Whole-Premise Baseline Calculation Methods*

Baseline calculation methods based on whole-premise interval metering data can be described in terms of three fundamental components.

- A set of data selection criteria,
- An estimation method, and sometimes
- An adjustment method.

These three components are explained below. Details on different procedures that have been used for each of them are then described.

#### *Data Selection Criteria*

The data selection criteria determine what data will be used in the baseline calculation. What data selection rules are appropriate depends in part on the estimation method. The selection criteria address both what should be included and what should be excluded. Inclusion is frequently determined by time period (e.g., a certain number of previous days). The other common approach is to select days that are similar to the curtailment day in terms of weather variables and/or system load. Exclusions are less varied across different calculation methods. All baseline calculation methods examined exclude hours in which there was a curtailment in effect. Other exclusion criteria are targeted to identifying abnormal load unrelated to curtailment events. These exclusions vary with particular methods.

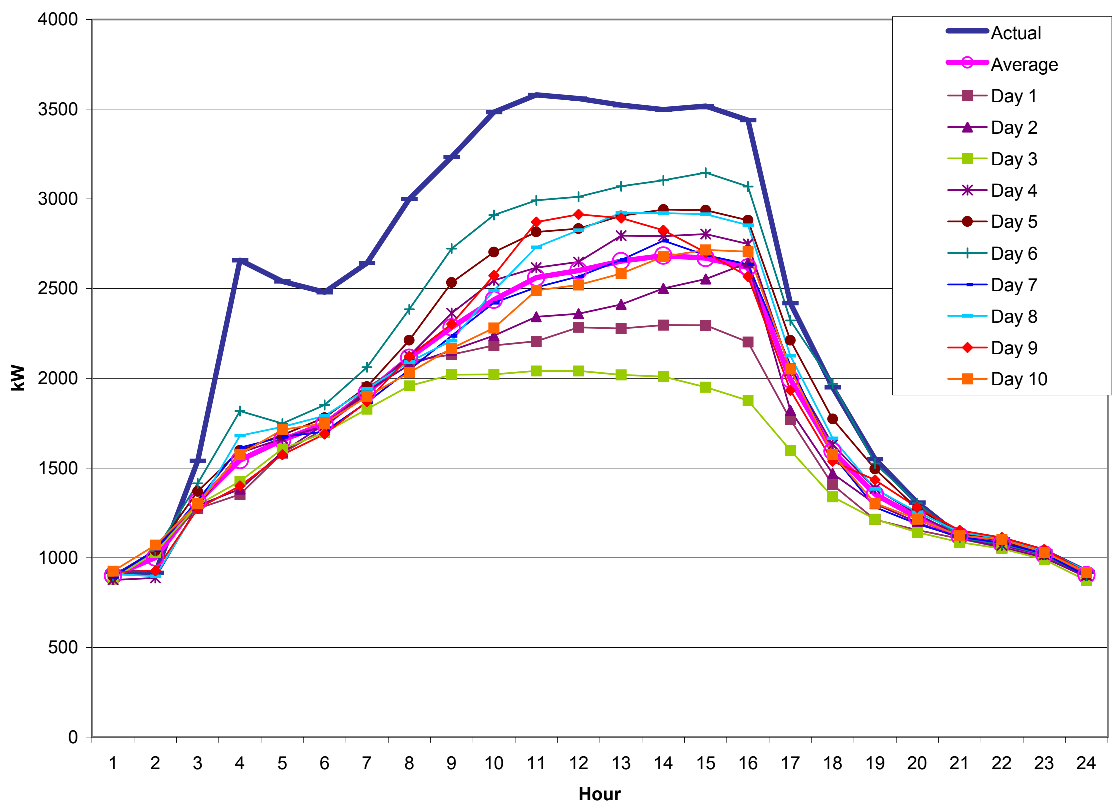
#### *Basic Estimation Method*

Once the data selection criteria have determined the appropriate data for a baseline calculation, the **estimation method** delivers an estimate of the load shape for the hours in question. If the selection criteria limit the data to a single load observation for each hour, there is no estimation method required. When the selection criteria return more than one possible load for each hour, the estimation method combines these data.

The two general types of estimation methods used are averaging and regression models. Averaging means that for each hour of the control period, the baseline is calculated as the average over all selected days of the load at that hour.

Figure 2–2 below illustrates a baseline calculated as the average of the past 10 uncurtailed business days. This is the baseline method used for the figure above. For the illustration below, a customer that was not curtailed is shown, so that the calculated baseline can be compared with the actual load. The figure shows the hourly loads for each of the 10 days, and the average. Also shown in the figure is the actual load curve on the curtailment day. The loads on the recent uncurtailed days are lower than the loads on the curtailment day. This is a common occurrence when the whole premise consumption, as well as the possibility of curtailment, is weather sensitive (e.g., curtailment is much more likely on a very hot day). As a result, the average understates the curtailment-day load.

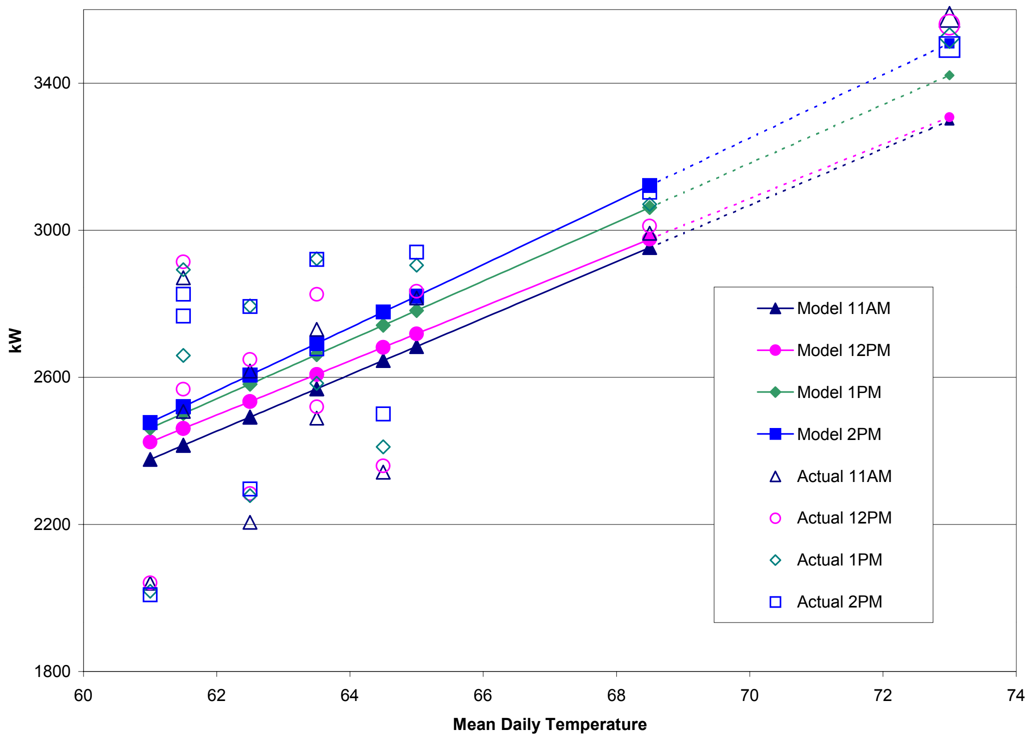
**Figure 2-2**
**Illustration of Baseline Calculated by Averaging**




A simple regression model using 10 days of data is illustrated in Figure 2–3. In this model, the load $L_{dh}$ at each hour $h$ of a given day $d$ is modeled as a function of the daily cooling degree-days $CDD_d$:

$$L_{dh} = \alpha_h + \gamma_h CDD_d.$$

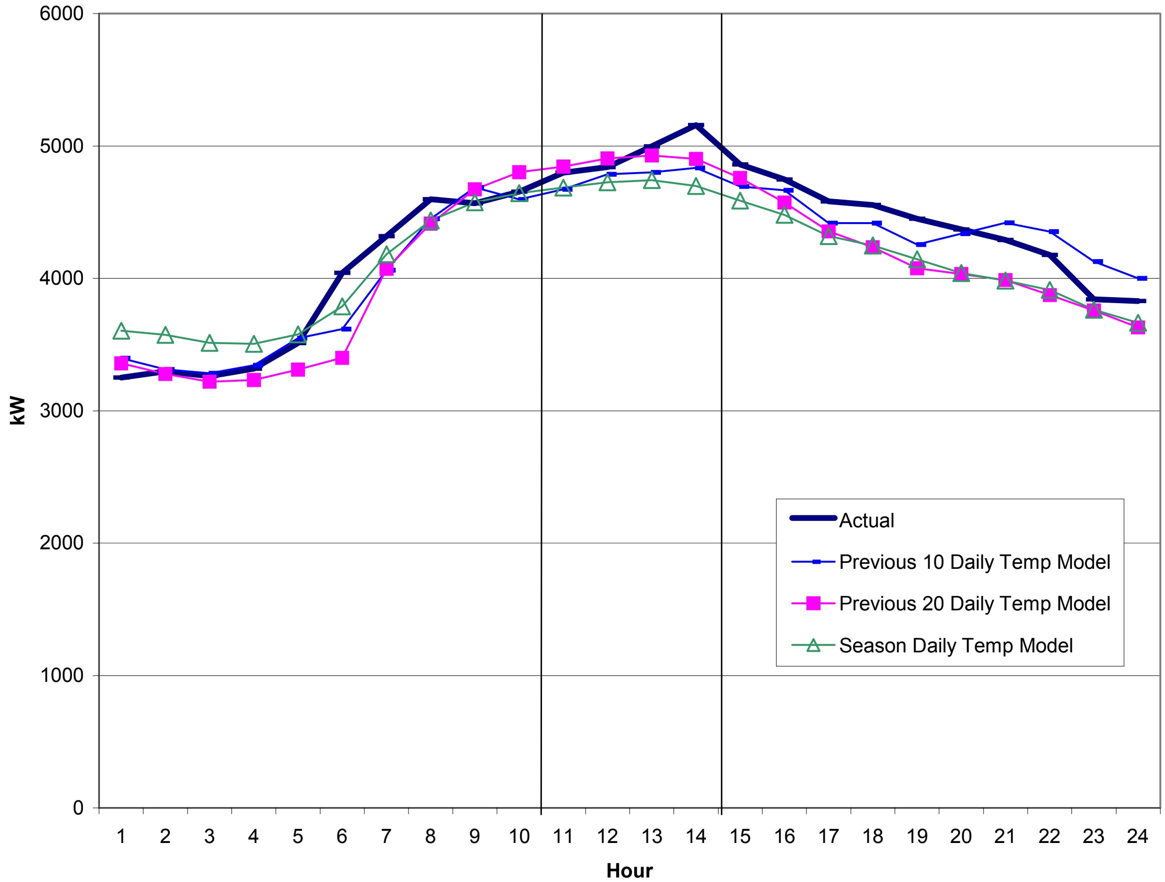
**Figure 2-3**
**Illustration of an Hourly Regression Model**




That is, for each hour *h* of the curtailment period, a separate line is fit to the loads at that hour. Figure 2–3 shows the actual load observations for the 10 days, for each hour *h* in the curtailment period. Also shown are the fitted lines for each hour. Days used in the regression all had daily temperature below 70°F. At right side of the plot, at a higher temperature than on any of the days used to fit the regression model, are the actual loads on the curtailment day. Despite the scatter in the data, the model does reasonably well at estimating the loads for this day. As in the previous figure, the illustration is for a day when curtailment did not occur, so that the actual load can be compared with the model.

The baseline estimated by this regression model for the curtailment day is shown in Figure 2–4. Also shown in Figure 2–4 are baselines calculated using data from the past 20 uncurtailed days, and using a full season of data. In this example, the models based on different lengths of input data give fairly similar results.

**Figure 2-4**
**Illustration of Baselines Calculated by Regression Models**
**Using Varying Data Selection Criteria**




### *Adjustment Method*

Many baseline calculation methods include an **adjustment method** to take advantage of data from the day of a program implementation. That is, a "provisional" baseline is adjusted to the actual load data of that day. Actual load from some hour or hours immediately preceding[1] curtailment is compared to the provisional baseline created by the estimation method. The provisional baseline is then adjusted to line up with the actual load for those last non-curtailment hours. Adjustment methods vary both with respect to which hours are used for the adjustment and the type of adjustment, typically additive or multiplicative. Further details and specific examples are described below.

[1] ERCOT considered employing an adjustment of load both preceding and following the curtailment call.

### 2.2.3 Further Details on Components of Baseline Calculation Approaches

#### *Common Combinations of Data Selection Criteria and Estimation Methods*

Baseline methods can be defined by any combination of data selection criteria, estimation method, and adjustment method. However, certain data selection criteria and estimation methods tend to be used together, because one is designed to compensate for limitations of the other. The common combinations are the following:

- **Single point data selection criteria – No estimation method required:** These methods involve the selection of a single day or single load levels as the baseline. As a result, no estimation method is required to combine data from multiple days. One example is a flat baseline defined by the load in the hour before curtailment, as in the PJM emergency program. Another is the selection of a single "best-fit" match day based on similarity of temperature, system peak load, or other criteria; in this case, the baseline for each hour is the load on the match day at that hour. Alternatively, the previous day's hourly loads may be taken as the baseline loads.
- **Multiple day selection criteria – Estimation method required for aggregation:** Multiple days are selected from the recent days. Possible selection criteria are described below. When multiple days are selected, some means is needed to aggregate the load data from these days into a baseline. Averaging by hour of the day is the most common method of aggregation. Regression models are also used.
- **Seasonal data selection criteria – Regression approaches:** When a full season of data are used as the basis for calculating the baseline, regression models rather than simple averages are typically used. Regression techniques allow for the explicit incorporation of additional variables other than loads on uncontrolled days. These methods provide estimates for a particular set of conditions, without requiring that the input data be screened to match those conditions. When a full season of data are used, regression rather than averaging is typically used, to develop an estimate for conditions similar to the curtailment day. However, regression methods can be used with shorter than seasonal data selection.

#### *Estimation by Averaging*

The most common approach is a combination of multiple day selection criteria and a simple average for each hour of the curtailment. Averaging has the advantage of simplicity but depends on the assumption that the selected data approximate what the curtailment day load would have been without the curtailment. If the selected days are expected to be similar to what the curtailment day would have been, the average based on a larger pool of days will be a more accurate baseline. On the other hand, increasing the number of days by including days that are less likely to be like what the curtailment day would have been can make the average less accurate as a baseline.

Key factors that can affect how similar the days in the baseline are to the curtailment day include daytype, weather, and structural/operational changes at the facility. Daytype is a classification of days according to the calendar, such as by day of the week (at a minimum, weekday or weekend), holiday, and season or month. Facility operations tend to vary with daytype. Weather affects heating and cooling loads. Structural or operational changes refer to either changes in the equipment or the building itself, or to long-term changes in operations such as changing shifts or schedules.

### *Data Selection Criteria Used with Averaging*

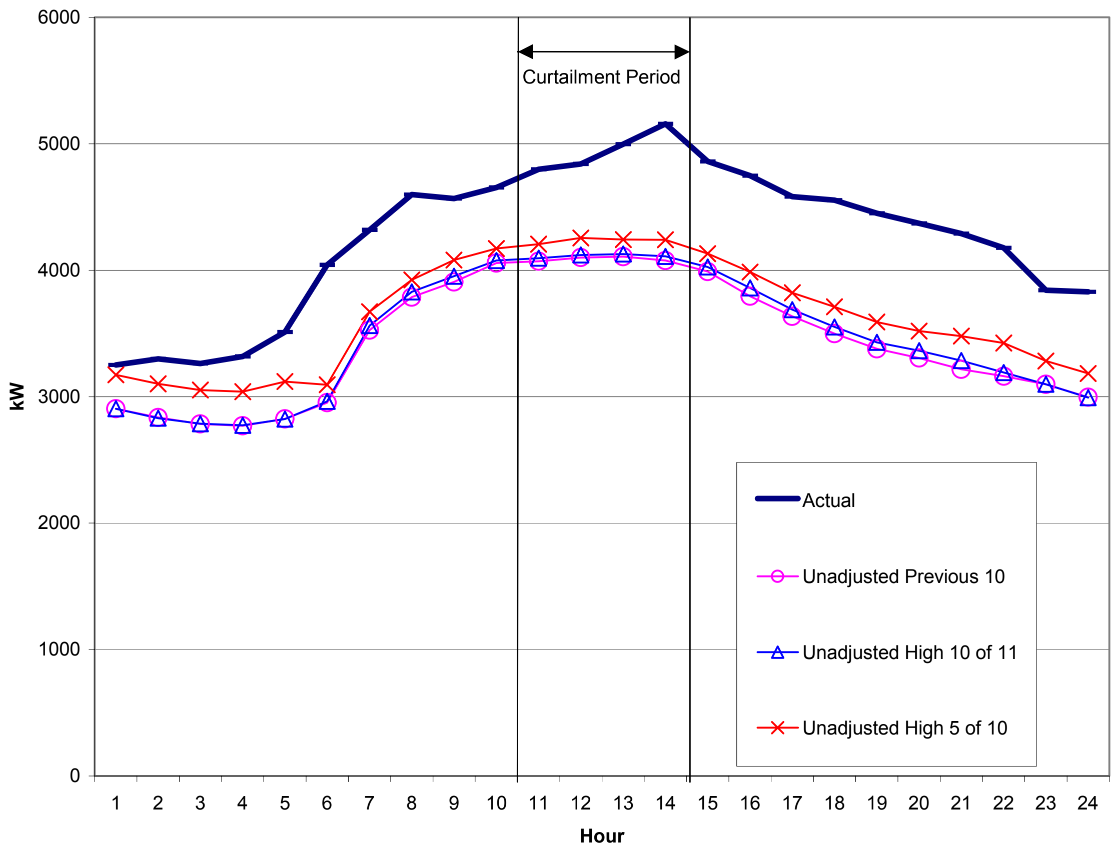
Data selection criteria used with averaging generally limit data so as to limit the potential for distortions in the baseline due to such factors. Most baseline calculation methods used recently by the ISOs limit data to the 10 or 11 preceding business days, excluding days when curtailments occurred. This data restriction removes weekend days and holidays and puts a limit on the ongoing effect of large-scale weather and structural changes. For non-weather-sensitive accounts with a consistent load shape within any two-week period this approach is suitable. Where weather and structural changes are clearly factors within a two-week span, these data selection criteria are limited in their ability to control for these factors in the averages.

Weather is a particularly critical issue when averaging is used and load is weather-sensitive. To the extent that the previous two weeks' weather is not similar to the curtailment day weather, an estimate based on averages will tend to understate baseline loads and corresponding demand reductions. The reason is that curtailment events are most likely to be called during periods of high load, which tend to correspond to extreme weather. Recent uncontrolled days will usually correspond to less extreme weather and loads. Furthermore, the exclusion of recent days with program implementation assures that representative load data will stay out of the sample regardless of whether curtailment took place at a particular site. Thus, estimated baselines for weather-sensitive accounts are likely to be systematically understated for the periods when the baseline calculation is most likely to be needed.

The understatement of load due to weather sensitivity can be addressed both indirectly and directly within the averaging estimation method. An indirect adjustment is to select for the baseline calculation only the recent days with the highest loads. An example is the "5 out of 10" rule used by the NYISO and PJM. From the previous 10 uncontrolled business days, the five days with the greatest curtailment period total load are selected. This approach reduces the problem of understatement of the baseline, but does not eliminate it. Even the highest 5 of the last 10 uncontrolled days will typically have lower loads than would have occurred on a day when curtailment was needed. That is, days when curtailments are called are likely to be more extreme than the days used to construct the baseline, when no curtailment was called.

Examples of alternative selection rules are illustrated in Figure 2-5 for a particular account on a curtailment day. This example is for an account that was not curtailed, so that the baseline estimates can be compared with the customer's actual metered load. This actual load is the heavy curve in the picture. The lowest curve is the baseline calculated as the simple average of the load at each hour, over the previous 10 uncurtailed business days. The dotted line is for the average of the highest 5 of the last 10 days. Between these two, just slightly above the simple average of the last 10, is the average of the highest 10 of the last 11 days.

**Figure 2-5**
**Examples of Alternative Selection Rules**




A more direct approach to selecting days for the baseline that are more similar to the curtailed day is to account for weather in the data selection process. Match day criteria, whether load or temperature based, can be used to select days with similar load characteristics. Clearly this approach only works to the extent that such comparison days exist. Match day approaches consequently require a sufficiently large pool of days to select from that comparable days are likely to be found. Match day criteria will be ineffective during the first heat wave of a summer unless either previous summer load is available for consideration, or the baseline calculation can be deferred until more summer data are available. However, any extension of the overall time span from which data can be drawn increases the possibility of structural changes affecting the accuracy of the baseline. This is less of a problem if current season data are used than if prior season data are used. However, for programs that involve financial settlement for individual control events, parties are typically reluctant to wait until the end of the season to determine curtailment amounts.

### *Estimation by Regression*

The other common estimation methods are regression-based approaches. An advantage of a regression estimation method is the explicit modeling of factors that cannot be accounted for in an averaging approach.

In simple terms, a regression model determines the relationship between some observable factors and load. Using this relationship, load can be estimated for any scenario based on the levels of those factors for that scenario.

Factors affecting a customer's load at a given hour include:

- Operating schedules across days and hours
- Activity levels, such as production levels at a manufacturing plant, or occupancy levels in a hotel or hospital
- Weather conditions that drive cooling and/or heating loads.

Load models tend to include terms for daytype, hour of the day, and weather. While activity levels are important drivers of load, it is difficult to obtain measures of activity that are objective and meaningful for a specific customer. In a customized model, activity measures appropriate to a particular customer can be incorporated.

Regression approaches also have disadvantages. Regression estimation is a more complicated technique than simple averaging. Ease of calculation and transparency are both important considerations for baseline protocols. Regression approaches also can require more data. The concerns with respect to longer data spans and structural load changes remain. That is, obtaining good estimates of the effects of the regression variables requires that data be available over a range of conditions; however, expanding the range of input days increases the chance of including days when the relationships being estimated in the model were different than on the curtailed day. Regression approaches operate on the assumption that the relationship between, say, cooling degree-days and load is the same on curtailment days as on the noncurtailment days that contributed to the regression estimates.

Finally, if the regression model includes factors that do not influence the particular customer's load, the model may simply add noise compared to a simpler average. In particular, using weather models for non-weather-sensitive accounts may do more harm than good.

### *Data Selection Criteria for Regression Models*

When regression models are used, data selection criteria are typically fairly simple. The key question is how far back, and possibly forward, to go in time. If the regression model is offered as an alternative to a default averaging method, the regression may use the same data selection rules as for the average.

Regressions based on the past 10 uncurtailed days have the advantage of using operating conditions likely to be similar, apart from weather, to those of the curtailment day. The disadvantage is that the short series means that the model is estimated less accurately. This is particularly a problem if the range of weather variation is slight, which makes it hard to estimate the effects of changes in weather. Moreover, with more limited data, the model diagnostics will be less reliable indicators of whether the model form itself is appropriate to this customer. If the model is not appropriate, using the fitted model to extrapolate to more extreme weather conditions, which is the typical application in this context, can introduce large errors. In addition, a smaller number of input days increases the potential effect of an occasional anomalous day.

Regressions based on a full season of data will tend to be more accurate. However, most program operators and participants do not want to wait until the end of the season to determine curtailment credits. Models based on the previous season can be used. This approach introduces errors if the operations have changed in the interim.

### *Adjustment Methods*

Adjustment approaches in use include:

- None (unadjusted).
- Additive adjustment. Adjusts a provisional baseline. A constant is added to the provisional baseline load for each hour of the curtailment period. For simple additive adjustment, the constant is calculated as the *difference* between the actual load and the provisional baseline load for some period prior to the curtailment.
- Scalar adjustment. Adjusts a provisional baseline. The provisional baseline load for each hour of the curtailment period is multiplied by a fixed scalar. For simple scalar adjustment, the scalar multiplier is calculated as the *ratio* of the actual load to the provisional baseline load for some period prior to the curtailment.
- Weather-based adjustment. A model of load as a function of some weather parameter is fit to historical load data. The fitted model is used to estimate load (a) for the weather conditions of the days included in the provisional baseline, and (b) for the weather conditions of the curtailment day. The difference or ratio of these two estimates is calculated, and applied to the provisional baseline as an additive or scalar adjustment.

#### **Additive and Scalar Adjustments**

The additive and ratio or scalar adjustments are illustrated in Figure 2–6. The unadjusted baseline is the average of the last 10 business days. The additive adjustment shifts this unadjusted baseline up to match the load in the two hour before curtailment. The scalar adjustment scales the unadjusted baseline to match the load in the same two hours. Both these adjustments bring the unadjusted baseline closer to the actual load during the curtailment period.

**Figure 2-6**
**Additive and Scalar Adjustments**
**to the Two Hours Prior to Curtailment**

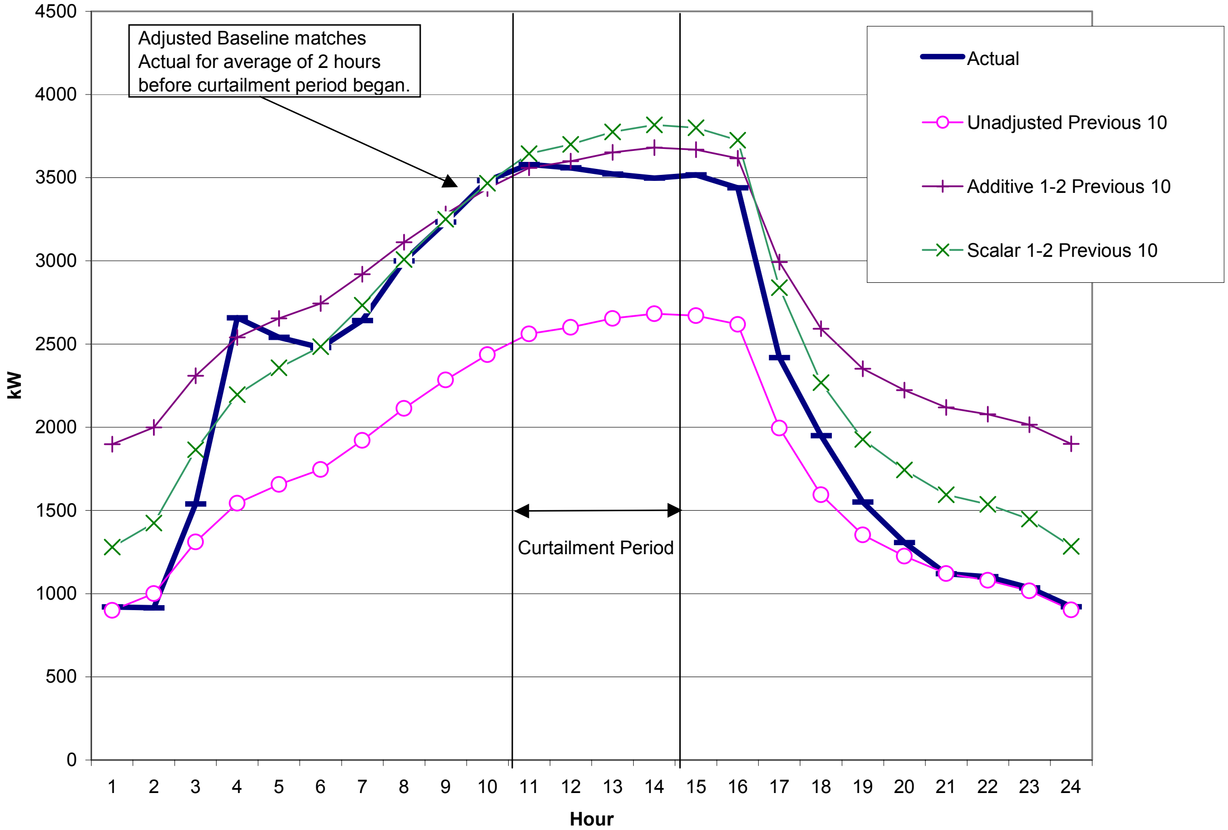
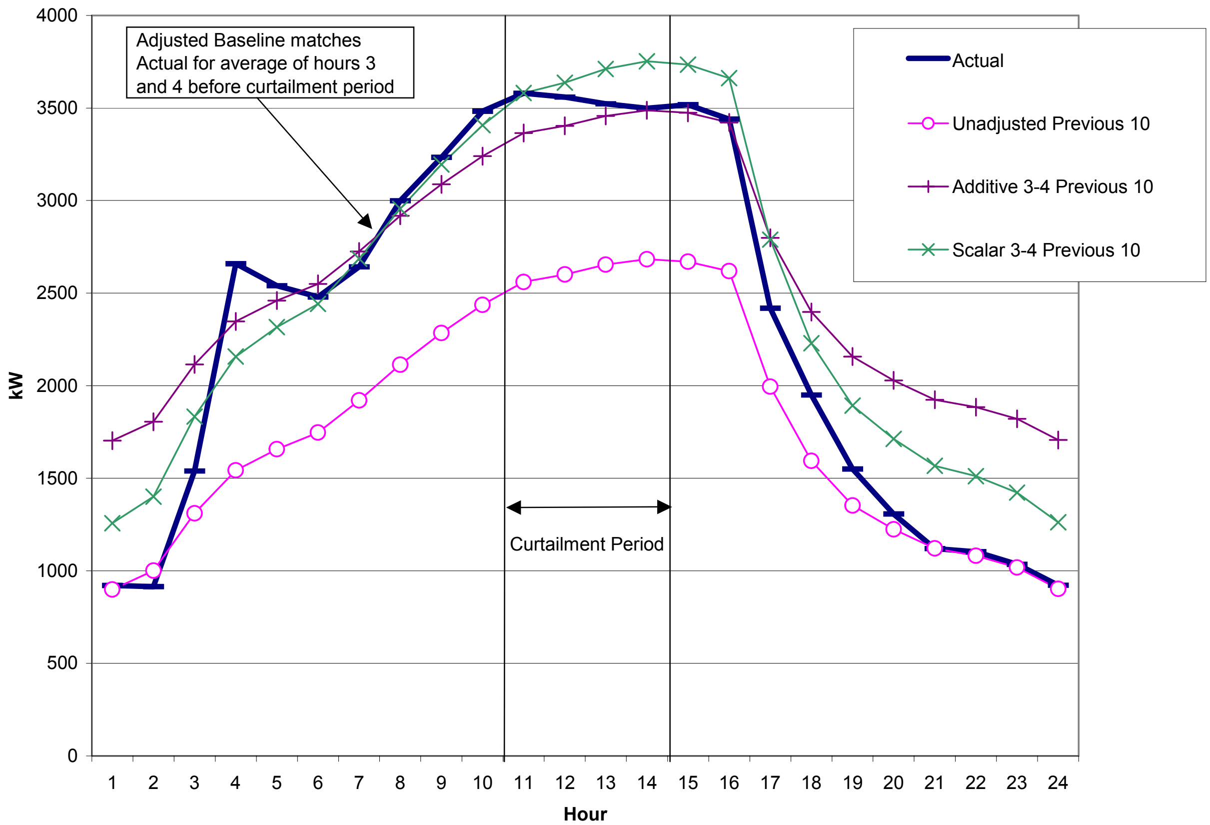



Figure 2–7 illustrates the same type of adjustments, but adjusting to the third and fourth hours before curtailment. This period is typically prior to the time customers would have been notified of the curtailment.

**Figure 2-7**
**Additive and Scalar Adjustments**
**to the Third and Fourth Hours Prior to Curtailment**




Reasons to prefer adjustment to earlier hours include the following:

1. There is less opportunity for the customer to manipulate the baseline by artificially boosting load after receiving a curtailment call but before the beginning of the curtailment period.
2. There is less possibility that the baseline will be set too low because the customer began curtailment promptly after notification, prior to the start of the formal curtailment requirement. In these cases, the adjustment to the two hours prior to the curtailment period would shift the baseline down to match already curtailed load.

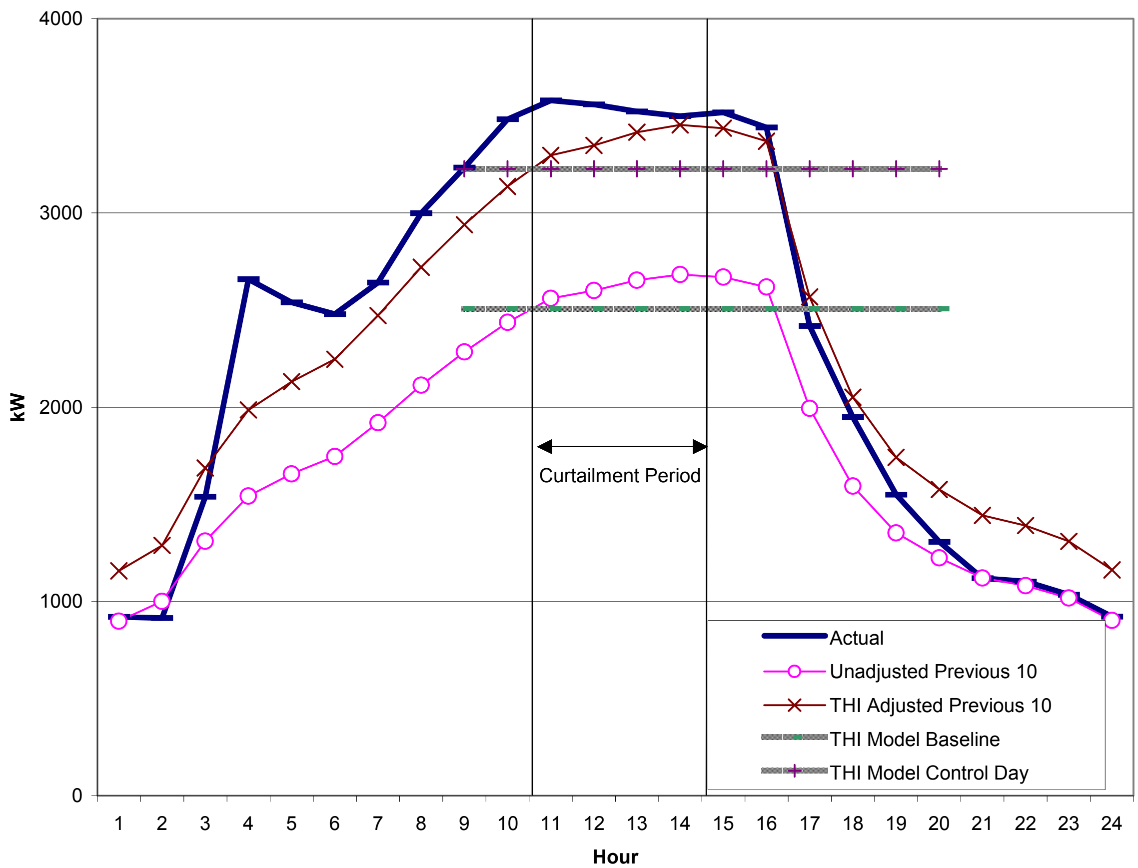
On the other hand, if neither of these reasons is a concern, the third and fourth hour prior to curtailment may not be as accurate a basis for adjusting the load as the closer first and second hours.

### PJM Weather-Based Adjustment

A weather-based adjustment is used by PJM. This method first fits a regression model to load as a function of the temperature-humidity index (THI). The time period used to fit this regression is longer than the limited window used for the provisional baseline. The fitted model is not used directly to construct the baseline. Instead, the model is used to estimate the load for two conditions, one the average THI for the peak hours of the curtailment day, the other the average THI for the peak hours on the baseline days. The ratio of the modeled curtailment-day load to the modeled baseline-day load is then used as a multiplicative adjustment to the provisional baseline based on a simple average.

The method is illustrated in Figure 2–8. The two horizontal lines spanning the peak hours indicate the THI model estimate of load for the curtailment day and for the baseline days. The bottom load curve is the provisional baseline, the simple average of the last 10 uncurtailed business days. The curtailment-day THI model estimate is around the average value of the actual load across the peak hours on that day, indicating that the THI model itself is doing a reasonable job of matching actual loads. Likewise, the baseline-period THI model estimate is around the average value across the peak hours of the provisional baseline. The final, adjusted baseline is the provisional baseline scaled by the ratio of the two THI model estimates. The THI-adjusted baseline is much closer to the actual load than is the unadjusted baseline.

**Figure 2-8**
**Illustration of PJM's THI-model Adjustment**



Use of a longer time period as input to the THI regression avoids some potential sources of modeling error. Use of the model to develop an adjustment factor mitigates systematic errors that may exist in the fitted model. Applying the adjustment factor to a baseline developed from recent data limits problems related to operational changes over the time span of the data. The same type of approach could be used with a different weather model, or with an additive rather than scalar adjustment.

### Effects of Adjustment

Adjustments may be made to any of the base estimation methods. Adjustment factors take advantage of very recent data to overcome potential limitations of any provisional baseline calculation. Adjustment factors have the obvious advantage of utilizing data from immediately before (and in some cases also after) the curtailment period.

Adjustment methods retain the daily shape produced by the estimation method, but translate it to align with known conditions of the curtailment day. An additive adjustment shifts the curve up or down by a constant amount. A scalar or ratio adjustment scales the shape by a constant amount.

Any such adjustment approach assumes the adjustment hours' load is a fair indicator of that day's load in the absence of curtailment. If the adjustment hours' load is abnormally high, the adjustment and thus the baseline will be high for the whole day. This high load could be the result of the perfectly sound policy of pre-cooling with the intent of shedding load during a curtailment period. It could also be a result of a strategic decision to increase the adjustment and thus increase the credited reduction (i.e., game the baseline). In fact, these two approaches would be indistinguishable to the outside observer and yet both would unfairly overstate the baseline.

Conversely, industrial processes frequently take time to shut down, or schedule operations in blocks of hours. Thus, if a curtailment is required to start at a particular time, it may be necessary to begin the curtailment substantially before that time. As a result, adjustment to the hours just prior to the official start of the curtailment period can produce an unfairly low baseline.

Finally, the hours used for calculating the adjustment amounts often occur during the daily ramping up period (e.g., 8–10 AM). There is potential for more variation during this period of increasing load. This effect can result in wide swings if adjustments are used.

Any of these issues can change an adjustment designed to refine a provisional baseline into an additional source of error. The volatility of adjustment period load, for any of the above reasons, is a fundamental concern with any adjustment approach.

### Additive Versus Scalar Adjustments

The two different adjustment methods, additive and ratio, assume a different relationship between adjustment period loads and curtailment period loads. The additive adjustment implicitly assumes that the observed difference between curtailment-day and baseline load at the adjustment hours reflects a difference in operations that would be the same in magnitude across all hours of the day. An example might be a constant-load process that is on or off for the day.

For weather-sensitive accounts, the difference between the provisional baseline and the curtailment day would generally not be expected to be the same in each hour. Nonetheless, several programs use additive adjustments as a rough correction for weather sensitivity.

Some practitioners believe that a ratio or scalar adjustment provides a better correction for weather sensitivity. This adjustment assumes that the difference between the curtailment-day load and the provisional baseline reflects factors that will vary in magnitude roughly in proportion to the hourly provisional baseline. If load is high in the early morning because the day is hot, the increased load will be even greater later in the day. There can also be cases where fluctuations in non-weather-sensitive accounts are better described by scaling a basic load shape than by an additive shift. An additional advantage of scaling compared to an additive adjustment is that scaling avoids the possibility of calculating negative loads for the adjusted baseline.

One way to view a regression approach is that it attempts to find an appropriate mix between adjustments for fixed loads and adjustments for variable loads. Even so, some methods use a mix of regression and additive or scalar adjustment to the curtailment day. One example is the PJM weather-based adjustment, described above. This method uses a weather model as the basis for adjusting a simple average provisional baseline. Conversely, a provisional baseline based on a weather model can be combined with a simple adjustment method, such as the additive adjustment to align with load in the two hours before curtailment.

### *Methods for Non-Weather-Sensitive Accounts*

While much of the attention in baseline method development has been given to addressing weather sensitivity, many of the customers participating in demand response programs may have little weather sensitivity. Indeed, many of the simpler methods that do not account for weather variation were developed in the context of programs with large, non-weather-sensitive accounts. For such loads, methods designed to mitigate weather-related biases can simply increase random errors.

Many programs address this issue by making weather-adjustment procedures optional. Customers enrolling in the program choose the form of the baseline calculation that will be used. For instance, PJM has three choices: (1) for non-weather-sensitive accounts, (2) for weather-sensitive accounts, and (3) the possibility of a custom baseline (that must be approved).

# 3 EXISTING METHODS

In this section, a number of existing demand response calculation methods are reviewed. This review was based on several sources:

- public documentation of baseline methods, in particular from ISO and utility websites
- papers shared by analysts who have developed methods, primarily from utilities
- interviews with analysts and stakeholders involved with the method development process at ISOs and utilities.

We begin by describing the process of developing the methods, and the key issues that were considered by the parties involved. We then describe a number of specific methods, and summarize their advantages and disadvantages.

## 3.1 DEVELOPING DEMAND RESPONSE CALCULATION METHODS

The demand reduction is calculated as the difference between the metered load and an estimated "baseline" in each increment. Thus, the central question is how the baseline is determined. As noted, the accuracy of the calculated baseline as an estimate of what the customer's load would have been in the absence of the program is only one of several criteria for defining the baseline calculation methods. Other criteria relate to the practicality of the method for operators and customers, and its potential effect on decision-making.

To better understand these other criteria, XENERGY interviewed 11 key participants in the development of the baseline methodology for five independent system operators (ISOs) and three utility programs. Notes of working groups, program and tariff filings reports, and program documentation also informed analysis of the baseline method development. We have also drawn on conference presentations and informal conversations with demand response program analysts and stakeholders in the baseline method development process.

### *3.1.1 Key Issues*

Key observations on the baseline development process based on this review include the following:

1. The baselines for ISO programs tend to be created by working groups exclusively focused on demand response programs, while small teams or individuals created the baselines of utility programs.
2. Baselines used by utilities for program evaluation are more likely to be calculated from a more sophisticated model.

3. The ISO baselines evolved between the 2001 and 2002 program years.
4. ISOs considered and used different portions of baseline methodologies of other ISOs, but each ISO had its own variant.
5. A baseline approved by another ISO was a significant seal of approval when considering various methodologies.
6. The ISO New England (ISO-NE) baseline development process was much less interactive than the other ISOs.
7. All parties understood that there were imperfections and tradeoffs in their baselines.
8. NYISO (New York Independent System Operator) and PJM (Pennsylvania Jersey Maryland Interconnection) working groups were many times contentious and political. The baseline was just one of many issues in implementing their programs.
9. Trade-offs were made between simple baselines, which were perceived to be less likely to capture the "real" load reductions absent the demand response program, and the more custom baselines for individual sites.
10. Gaming (manipulating load to create a more advantageous baseline) and/or free-ridership (obtaining incentive payments for load reductions that would have taken place without the program) were a major concern in baseline development.
11. There are sharp differences on the issue of demand response and shut-downs. Should all day / long-term shut-downs be compensated? Most of the baselines will not compensate for multiple-day shut-downs.
12. In some cases (the 2001 California ISO Discretionary Load Curtailment Program, and the 2001 and 2002 PJM programs) the managing organization allowed the participant to propose an alternate measurement program that would be reviewed and approved by the ISO. For the California ISO (CA ISO), this allowed for other options such as engineering analysis, operational or test data, use of loggers versus interval metering, etc.
13. While there have been many discussions of the implications of choosing different types of baselines, there has been little statistical justification of choosing a particular baseline. They are all works in progress.

These observations substantially mirror the experiences of developing the energy efficiency IPMVP. This activity involves highly technical situations and requires trained professionals for some of the engineering and statistical calculations. At the same time, establishment of a savings calculation method is primarily a contractual undertaking that seeks to apportion risk and provide fair rewards. Thus, an important first step is to create a framework that any sensible person could understand. Details and refinements can be added in time. Minimizing the number of different variations of reasonable approaches can improve the attractiveness of programs and increase participation levels.

### *3.1.2 The Process of Developing the Methods*

From a process point of view there is generally a sharp differentiation between the utility and ISO development of baselines for the purposes of measuring the impacts of demand response programs. For the ISO programs, the purpose of the baseline is to determine payments to participating customers for individual curtailment events. For traditional utility programs, payments to customers are often not tied to explicit calculations of load reductions.

Even where new utility programs are designed to create market-based mechanisms for demand response, the utility is generally able to develop its baseline approach largely at its own discretion. Their program designs are subject to regulatory review, but are generally not the subject of wide public debate.

By contrast, the ISO, which acts as an impartial market operator, tends to develop baselines through a participatory stakeholder process. The ISO baselines generally have been created via committee, where many parties give their input. The baseline is only one component in the design of a demand-response a program. These programs must be approved by vote. One respondent noted that simplicity in explaining the baseline methodology to the voting members was an important criterion.

In addition at the ISO level, demand response programs are looked at warily from entrenched market participants who may lose revenue or be assessed program costs. Counterbalancing this resistance is pressure from the Federal Energy Regulatory Commission (FERC), and state Public Utility Commissions (PUCs) to make economic demand response an important integrated part of the market.

The implementation of the PJM economic load response program has been particularly contentious. Numerous comments and protests have been filed at the FERC for the proposed 2002 program (Docket #ER02-1326). The baseline methodology, and even the use of a baseline as a program component, has been called into question. Specifically in regards to the baseline, PJM has had to defend the use of the five highest days out of ten methodology and the weather-sensitivity adjustment. These method elements are discussed in Section 4.

### *3.1.3 Desirable Features in a Baseline Calculation Method*

Most of the goals developers described for the baseline were similar. They were

- to reflect the estimated level of load (kW) that would have occurred in the absence of the program,
- ease of use for program participants,
- ease of use for program administrators, and
- deterrence of gaming.

Given the multiple objectives, all baseline developers understood that the baseline methodology they chose was a compromise. Development of the ISO's baselines were further complicated by the political realities of the approval process. None of the respondents believes their method is the universal best way to calculate the baseline. In practice, the baseline methodology has been driven by the comparative importance of the objectives of the demand response program, the wholesale environment, the structure of the group choosing the baseline, and financial considerations. The financial impact includes cost to implement (e.g., need for sub-metering) and assessment of program costs (e.g., who is financially responsible for incentive payments).

In program documentation, criteria for the baseline choice were not explicit, with an exception of the development of the Electric Reliability Council of Texas (ERCOT) baseline. Criteria reported by interview respondents generally included:

- Simplicity
- Ease of use
- Ease of understanding
- Verifiability
- Accuracy
- Lack of bias (i.e., no systematic tendency to over- or under-state reductions)
- Ability to handle weather-sensitive accounts fairly
- Minimization of gaming
- Ability to be known prior to customer's commitment to a particular curtailment amount and event
- Costs for participant and operator to implement
- Consistency with other ISOs[1]

While not necessarily mentioned explicitly, simplicity, ease of use, and ease of understanding are all related to a desire to minimize the burden on participants and program operators. Burden means costs, as well as related time and nuisance factors.

Accuracy has two aspects. One is lack of bias. Bias is a systematic tendency to over- or under-state the baseline and the corresponding demand reduction. The second aspect of accuracy is variability. A method may be close to unbiased, that is, to be close to correct on average, yet have a high variance, meaning it tends to have large errors in either direction. Methods that have high variance are unreliable, and add risk to program participation and operations.

[1] Indeed, PJM cites consistency with NYISO program as a positive program attribute in a filing to the FERC. *For example, PJM adopted a Customer Baseline Load ("CBL") calculation methodology that is very similar to the one used in the New York program and has thus standardized the measurement of load response.* Response of PJM Interconnection L.L.C. To Protests. Docket ER02-1326 April 22, 2002.

Handling weather-sensitive accounts fairly is a question of both limiting bias and limiting variance related to weather-sensitivity. At the same time, methods introduced to improve baselines for weather-sensitive accounts should not increase errors for non-weather-sensitive accounts.

Gaming issues and the value of knowing the baseline in advance are discussed below.

## 3.2 Contentious Issues in Baseline Development

As noted, there are a number of desirable features in baseline methods. Attempting to balance competing objectives has been a challenge to the groups that have developed the rules for different areas. Following is a discussion of some of the most contentious issues.

### *3.2.1 Data Selection and Weather Adjustment*

Two of the most contentious issues for the baseline has been the justification of the days used to calculate the baseline and the use of a weather-sensitive adjustment. These issues are related, in that the selection of days is used in part to address weather sensitivity. Both PJM and the NYISO have used a "high 5 of 10" method, which selects the 5 days with the highest loads from the past 10 business days. The use of a particular window of days such as the highest 5 of 10 is justified explicitly by PJM in its filing.

> *By eliminating the five lowest usage days, PJM creates a baseline that reflects load conditions that most approximate those conditions when load reductions are likely to occur — i.e., when it is hot and consumption is high. Under cool conditions (i.e., low usage days), participants are not likely to reduce load; therefore, to include the low usage days in the calculation would create a baseline that is not representative of the participant's load absent a load reduction. Further, contrary to Mirant, the baseline methodology should include a weather-sensitivity adjustment, which will serve to either raise or lower a customer's calculated baseline according to weather fluctuations, eliminating the possibility that utilization of the highest five days prior to a relatively cool day will result in over-estimation of the customer's actual load.*

The 5 highest out of 10 day methodology was debated at PJM and the NYISO. An alternative methodology of the lowest 5 days out of last 10 was proposed by Mirant at the PJM working group level (see http://www.pjm.com/committees/user_group/dsrwg/meeting/20020225_meeting_materials.html). Mirant contended that this method would be much more likely to deter gaming and reimbursement for load reduction not actually delivered. While this is probably correct, the methodology was not endorsed at the working group level as other priorities prevailed.

The "high 10 of 11," the average of the highest 10 of the last 11 uncurtailed business days, was employed by the California ISO for its 2001 Demand Relief Program. The adoption of this methodology was relatively uncontroversial according to a California ISO representative. In 2000 the California ISO had a smaller program that used the average of all 10 of the last 10 days for the baseline. This methodology was exactly the same as used by the major utilities for their (similar) programs, and was easily adopted for the 2000 program. In the design of the 2001 Demand Relief Program, the California ISO got input from various stakeholders and reached a consensus to use the previous 11-business days and drop out 1 day that was the lowest demand.

The New York ISO methodology evolved from 2001 to 2002. In 2001 there was the implicit assumption used that the use of the high 5 of 10 methodology was an approximate weather-sensitivity adjustment. According to interviewees, this proved to be insufficient. Accordingly, the New York ISO looked for ways to improve their baseline for weather sensitive accounts for the 2002 baseline. To this end they included an adjustment similar to that used by the ISO New England (ISO-NE).

While most ISOs now include an adjustment as a way of addressing weather-sensitive accounts, the specifics vary. The ISO-NE baseline uses an additive adjustment to the two hours prior to curtailment, as described in Section 2. That is, a constant amount is added to the initial baseline load in each time increment, so that the adjusted baseline matches the metered load for the average of the two hour prior to the curtailment period. The NYISO, ERCOT, and PJM use a scalar adjustment. That is, the initial baseline load in each interval is scaled by a constant factor. The PJM adjustment, as described in Section 2, does not align the baseline with the observed load at a previous hour; instead, the adjustment scales the provisional baseline according to the relationship between the curtailment day and the baseline as indicated by a weather model.

There has been little hard evidence offered to justify either the additive or the ratio adjustment process. These adjustment procedures are described further in Section 5, and their performance examined.

The utility programs use more sophisticated and more opaque, but possibly more accurate methods to calculate baselines. The Cinergy method is the prime example of this type of methodology. It is econometrically sophisticated (e.g., uses state space and Kalman filtering to estimate mean daily load in the first of two stages of baseline computation), and is proprietary. The Cinergy method almost certainly would be rejected as too complicated for an ISO method. The ISOs, because of their process (e.g. consensus or vote at working group and then approval votes at higher levels) were constrained to relatively simple, straightforward, methods that are (relatively) easy to understand, and replicate.

### *3.2.2 Gaming and Free Ridership*

Establishing rules to limit opportunities for "gaming" or "free ridership" has been a frequent concern in baseline method development. Gaming means manipulating loads deliberately to inflate the baseline artificially and obtain excess incentive payments. Free ridership means obtaining payments for actions that would have been taken anyway without the demand response program. Many data selection rules have been established to limit gaming and free rider opportunities. Some special situations are worth noting.

One situation of particular concern is plant shut-down. Other concerns relate to the potential for demand reductions to be credited when none has taken place. Sometimes procedures designed to reduce systematic over- or under-statement in one situation can generate an opposite potential error in another.

### Shut-Downs

A good deal of discussion in the creation of baselines has been the appropriate way to handle temporary complete or partial closure of customers sites resulting in drastic reductions in load (also known as "shut-downs"). For an industrial plant that shuts down for two weeks over the summer, there have been debates over whether the plant should be able to claim compensation for that load reduction. Many argue that if the plant was going to shut down regardless of a demand response program, and then obtains "windfall" payments from the demand response program, this is a case of free ridership. That is, payments that were intended to provide incentives to change customer behavior are being made to a customer whose behavior is the same as it would have been without the program.

One way the shutdown problem is addressed is by using a baseline with a scalar adjustment to the hours just before curtailment, as described in Section 2. With this adjustment, zero usage during the pre-curtailment period results in a baseline of zero during the curtailment period. Thus, there is no opportunity to collect windfall payments as a free rider in the demand response program for the shut-down.

On the other hand, shut-downs may be initiated, at least in part, because of program incentives. Screening out the largest responses may take away an appropriate mechanism to signal such behavior.

### Minimum Bids to Limit Free Riding on Weather or Shut-Downs

For some economic bidding programs (e.g., New York and PJM) only bids with a minimum price are accepted (e.g., $50–$75/MWh). This rule inhibits participants from bidding demand reduction that may be based only on weather effects.

For example, if the baseline is an average with the high 5 of 10 data selection rule, the baseline would tend to overstate expected load on a cool day for a weather-sensitive account. The customer could therefore bid a "reduction," and count on receiving credit for it, without actually doing anything to reduce load compared to normal operations.

Employing a minimum bid to inhibit free riding on weather implicitly assumes that high prices are strongly correlated with extreme weather conditions. If the minimum bid price is $75/MWh, and if the price of energy (generally) goes above $75/MWh only on hot days, then it is impossible to get a "free ride on weather" on cool days. That is, the customer could bid \$75/MWh on cool days, but the bid (generally) would not be accepted because the clearing price would be below that level.

In the absence of a day-of adjustment, the minimum price bid would also deter free riding on shut-downs. For example, if a plant was to be shut down for a week, it would be easy to bid a low price into an economic program for the week, and know with near certainty that the load reduction bid would be accepted. The minimum bid price would preclude the acceptance of such bid for "normal" days.

Of course, picking an appropriate minimum bid price is very important. If the market generally has prices higher than the minimum bid (e.g., \$75/MWh in the California ISO during early 2001), then free ridership potential would be high. Conversely, if the market minimum price is set too high, there will be few if any opportunities for demand response participation, and little incentive to invest in demand response systems and technology.

**Manipulating Baseline Loads**

Gaming, that is altering behavior to artificially inflate a baseline, was also a major concern for many developers of baselines. A number of programs (e.g., ISO-NE, PJM, NYISO) screen out large changes in load in the creation of the baseline. For example, the ISO-NE 2002 program excludes days with four or more consecutive hours that are less than 75 percent or greater than 125 percent the average of a provisional baseline. This means that if a customer sees a heat wave coming, maneuvers of using additional energy to increase their baseline will be restricted.

### *3.2.3 Advance Customer Knowledge of the Baseline*

Another factor considered was the customer's ability to know its baseline prior to committing to a load reduction, particularly for bidding programs. Knowing the baseline prior to commitment reduces customer risk. To meet a program commitment the customer has to reduce total load to the computed baseline level less the amount of the load reduction bid. If the baseline is not determined until after the load reduction commitment is made a risk-averse customer will commit less load reduction than they can deliver in order to ensure that they can deliver their full load-reduction commitment. Generally, customers will not know their baseline prior to load reduction commitment when the baseline includes a day-of curtailment adjustment(s). If a weather adjustment is used, then it is likely that the customer will not know their baseline until after the day (curtailment period) is over as the adjustment will be based on the day (curtailment period) weather conditions.

If the baseline is not determined until after the curtailment, the customer can't know until after the curtailment whether they will be credited with enough load reduction to fulfill their bid commitment. As many programs have penalties for non-performance, bidders will tend to be conservative in their bids.

Not knowing the baseline in advance also increases the cost of participation in these programs, as the bidders must forecast what they believe their baselines will be prior to bidding. That is if baseline is based on weather, then the customer will have to procure a weather forecast, and forecast a baseline in order to determine how much load they should actually reduce in order to meet their load reduction commitment.

On the other hand, knowing the baseline in advance could lead to gaming. For days when the baseline seems to be understated (e.g., cool days after a heat wave), customers could confidently bid load reduction, with the expectation that they will not have to alter their operations at all to be credited with the full load reduction.

Advance customer knowledge was a major issue in the development of the NYISO baseline. For this ISO's day-ahead bidding program the baseline is based on days at least two days prior to the event, or one day prior to the bid. This has the dual effect of

1. allowing the customer to know at the time of the bid decision the baseline that will determine the customer's incentive, and
2. making it more difficult to game the baseline.

For example, in the New York ISO day-ahead bidding program, a bid for Wednesday load reduction is made on Tuesday. Possible days in the baseline begin with Monday and work into the past. Thus, a participant cannot blatantly increase load on Tuesday to boost its baseline. This procedure does not preclude increasing load in the previous days to increase the baseline, but the uncertainty of the market-clearing price makes such behavior risky for participants contemplating premeditated gaming.

## 3.3 Characterization of Different Baselines

A categorization of baseline calculation methods reviewed is provided in Table 3-1. More detail on these protocols is provided in Appendix A. A brief discussion of the key features of the methods presented in the table follows.

# Table 3-1
## Summary Baseline Protocol Matrix

| | Load Type Differences Addressed* | How Weather Sensitivity is Addressed** | Time Interval | Data Selection: Initial Timeframe | Data Selection: Final Selection | Data Selection: Excluded Days (other than weekends and previous program Control days) | Estimation Method | Adjustment Hours, If Adjusted |
|---|---|---|---|---|---|---|---|---|
| PJM-Day Ahead 2002 | WS/NWS, Self Gen, Cust. Spec. | Top 5 of 10, Optional adjustment to control-day weather | Hourly | 10 days, n-2 to n-11 | High 5 of 10 | Low Output Days. | Interval Average | h-1,h-2. |
| PJM Emergency | None | None | Hourly | Hour before | Same | None | | |
| ISO-NE 2001-2002 | WS/NWS, Self Gen, | Adjustment to control-day load | Hourly | 10 days, n-1 to n-10 | Same | Extreme Output Days. | Interval Average | h-1,h-2. |
| NYISO-DADRP 2001 | Self Gen | Top 5 of 10 | Hourly | 10 days, n-2 to n-11 | High 5 of 10 | Low Output Days. | Interval Average | |
| NYISO-DADRP 2002 | WS/NWS, Self Gen | Top 5 of 10, Optional adjustment to control-day load | Hourly | 10 days, n-2 to n-11 | High 5 of 10 | Low Output Days. | Interval Average | h-3, h-4. |
| ERCOT-BUL 2002 | WS/NWS | Optional adjustment to control-day load | 15 minute | 10 days, n-1 to n-10 | Middle 8 | None | Interval Average | h-1, h-2. |
| CAISO 2001#1 | None | None | Hourly | 10 days | None | None | Interval Average | |
| CAISO 2001#2 | None | None | Hourly | 11 days | None | None | Interval Average | |
| XENERGY | WS/NWS | Regression-based estimate, Adjustment to control-day load | Hourly | Variable | None | None | Regression-based | h-1, h-2. |
| LBNL/Kinney | WS/NWS | Regression-based estimate | Hourly | 10 days, n-1 to n-10 | None | None | Regression-based | |
| Nexant | WS/NWS | Adjustment to control-day load | 15 minute | 10 days | None | None | Interval Average | h-1 |
| Utility A | WS/NWS | Adjustment to control-day load | Hourly | Previous Month | None | All Days that do not fit the match-day criteria. | Interval Average | one hour, 8am - 11am |
| Utility B | None | None | Hourly | 5 days | None | Customer-specificed anomalous loads | Interval Average | |
| Utility C | WS/NWS | Regression-based estimate | Hourly | Undefined minimum data | None | None | Regression-based | |
| Utility D | WS/NWS | Regression-based estimate | Hourly | Weekdays, June through September | None | None | Regression-based | 5am - 10am |
| Utility E | None | Match based data selection | 15 minute | Undefined | 10 Days with min. SSE compared to day n-1 | None | Interval Average | All match-day hours. |
| Utility F | WS/NWS, Cust. Spec. | Adjustment to control-day load | Hourly | 2-3 previous years | None | Anomalous loads | Interval Average | h-1, h-2. |
| CMTA Proposed OBMC | WS/NWS | Adjustment to control-day load | Hourly | 10 days, n-1 to n-10 | None | None | Interval Average | h-1 through h-4 |

* WS/NWS: Different methods for weather-sensitive and nonweather-sensitive loads
Self Gen: Different methods for onsite generation

** Top 5 of 10: Select 5 days with highest average load during the hours curtailed on the curtailment day

### 3.3.1 Purpose of the Demand Response Calculation

The methods reviewed for ISO programs have all been used as the basis for payments to participants for particular curtailment-day reductions. Many of the other methods reviewed were developed as proposals or custom approaches for use in the ISO programs. Most of the utility methods reviewed were developed for program evaluation rather than as a basis for payments.

### 3.3.2 Addressing Weather-Sensitive Accounts

Many of the procedures reviewed allow different calculation methods for weather-sensitive and non-weather-sensitive accounts. Methods used to address weather sensitivity include regression models with weather terms, additive and scalar adjustments to curtailment-day load before curtailment, data restriction to days with higher loads, and match days based on load or weather match criteria.

### 3.3.3 Time Increment

Most of the methods are defined for hourly load data. A few use 15-minute data. The same methods could be used with whatever time increment is required for settlement.

### 3.3.4 Data Selection Criteria

Most of the methods that use a version of averaging, with or without adjustment, use 10 or 11 noncurtailment business days prior to the curtailment day as the basis for the baseline calculation. Restriction to the 5 days with highest average load out of these 10 and exclusion of extreme high or low load days are common screens. Rules for which days may be included, how far back in time to go to replace excluded days, and the sequence for screening days out and replacing them vary widely across programs and are often complex. Details are provided in Appendix A.

Methods that rely on regression or match days often use a full season or even multiple years of data in the regression. A few regression methods, developed as alternative to ISO averaging methods, rely on the same 10 days as required for the average.

### 3.3.5 Estimation Method

All the methods reviewed can be characterized as some form of averaging or regression, sometimes with adjustments. The one exception is the PJM emergency program, which uses the single hour before control as the baseline for the entire curtailment period. Thus, there is no "estimation" required other than selecting this hour's load.

### 3.3.6 Adjustments

Most of the adjustments to curtailment-day conditions match to the load one to two hours before curtailment. Additive and scalar adjustments were roughly equally common.

## 3.4 Pros and Cons of Alternative Approaches

No baseline methodology is perfect. Thus implicit in the choice of any baseline methodology are tradeoffs. The table below summarizes some of the most prominent pros and cons of various baseline methodology components.

**Table 3-2**
**Pros and Cons of Key Baseline Methods and Combinations**

| Baseline Method | Variant | Pros | Cons |
|---|---|---|---|
| Average | Any | Simple, easy to use and understand, low cost | Tends to understate baseline for weather-sensitive loads. |
| | High 5 of last 10 days | Partial adjustment for weather-sensitive loads | Still tends to understate baseline for weather-sensitive loads |
| | | | Can allow windfall load reduction credit on cool days |
| Regression | Any | Provides baseline corresponding to particular weather conditions of curtailment day | More complex, harder to understand, higher cost |
| | | | If observations don't include conditions as extreme as the curtailment day, model estimate may be inaccurate |
| | | | If account isn't weather-sensitive, may be less accurate than simpler methods |
| | Full Season | Adequate data and range of variation to yield accurate coefficients | Operating conditions from the period data are taken from may be different from curtailment day |
| | Recent 10 days | Operating conditions more likely to be similar to curtailment day | Model based on limited data may be inaccurate |
| Adjustment to precurtailment hours | Any | Simple, easy to use and understand, low | May be potential for gaming behavior during day-of-curtailment adjustment period |
| | | Adjusts to weather and operating conditions of curtailment day | Appropriate pre-curtailment increase in load (e.g., pre-cooling) will result in overstated baseline |
| | | Limits potential for collecting windfall credits for planned shut-downs | Pre-curtailment decrease in load in response to curtailment request (e.g., long ramp-down, canceling a shift) will result in understated baseline |
| | Additive | May adjust well for load change that is constant throughout day (e.g., industrial processes) | May not be appropriate if load changes during curtailment period (ratio adjustment may be better suited) |
| | Scalar | May adjust well for load change that is function of exogenous factor throughout day (e.g., higher levels of occupancy) | May not be appropriate if the day-to-day load variation is constant over the day (additive adjustment may be better suited) |
| Weather-Based Adjustment | Any | Explicitly takes into account weather conditions | Adjustment may not be known to customer until after curtailment period (i.e., until after weather conditions are known for the day) |
| | | | If no observations are available for extreme conditions, estimates used for adjustment may be outside range of model |
| | | | Will badly predict load reductions if the buildings are dominated by internal loads |

# 4 TESTING APPROACH

A major element of this project was to test several alternative baseline methods on several data sets from different locations. Interval load data were provided for customers who have participated in some type of demand response or controllable rates program. In addition, for most locations load data were also provided for a group of customers of similar size who were not participants in such a program. Performance measures for different baseline methods were developed by analysis of both curtailed and uncurtailed accounts.

In this section, we describe

- the baseline calculation methods selected for testing
- the time periods studied for both controlled and uncontrolled accounts
- the statistics calculated as measures of method performance.

## 4.1 BASELINE CALCULATION METHODS TESTED

The taxonomy of existing methods presented in Section 3 classified methods based on three elements:

- data selection criteria
- estimation method
- adjustment method.

We tested various combinations of these elements, rather than testing only those combinations currently being used. Some of the selection criteria in use for specific methods are complex, with an iterative process of screening days out, then going back further in time to replace the eliminated days. For purposes of the tests, we simplified some of these screening rules. We do not believe these simplifications substantially alter the performance characteristics of the methods.

It was not practical to test all possible combinations of candidate selection criteria, estimation, and adjustment methods. We did try to test each component in combination with enough variations in the other components to get a good sense of the effect of each. In addition, we made sure to include the combinations that corresponded most closely to key methods currently in use. In particular, the methods tested include those in use by the ISOs in California, New York, New England, and PJM, except for some details on data selection criteria.

### 4.1.1 Data Selection Criteria

#### Initial Selection

The initial selection of days for the ISO programs all involve something like the last 2 weeks of business days. Methods used by utilities for program evaluation tend to involve data for a full season. Variations we considered are the following:

- previous 10 business days beginning on d0-1 (California and New England ISOs)
- previous 10 business days beginning on d0-2
- highest 10 of the last 11 business days, beginning d0-1
- highest 5 of the last 10 business days, beginning d0-2 (New York and PJM ISOs)
- previous 20 business days beginning d0-1
- 20 business days from d0-10 to d0+10.
- entire season that includes the control day
- entire season from the previous year that includes the control day

where day d0 indicates the curtailment day.

#### Exclusions

All existing methods delete from the data selection any days that had a control event. Some replace these excluded days by going farther back in time as needed. Similarly, some methods screen out days of low or extreme output based on varying criteria and with or without replacement. For this analysis all excluded control days were replaced with the next eligible day. No extreme output exclusions were tested. If extreme output exclusions were used they would likely affect high variability account results.

### 4.1.2 Estimation Method

The two broad types of estimation methods are averaging and regression. We describe these methods more specifically below. In the analysis conducted, all load data were available on an hourly basis, and the analysis was conducted on this basis. In some markets, load data are collected and demand response is calculated on a finer time interval, such as half-hour or 15-minute. Our discussion refers to analysis of hourly data. However, the same principles would apply with data at a finer time increment.

#### Averages

Averaging means that the baseline for each time interval of the curtailment day is calculated as the simple average, across all the days chosen by the data selection criteria, of the loads at that time interval. For example, the baseline for the hour ending 1 pm is the average over all the selected days of the loads on those days for the hour ending 1 pm.

### *Regression Models*

#### Types of Models Considered

Regression can take a wide variety of forms. It's beyond the scope of this work to attempt to determine the best general regression model for this application.

In the context of calculating demand response baselines, the regression model uses the data selected for a particular account and event. The model is fit to those data, and applied to the conditions during the event, to estimate the load that account would have in the absence of the control or curtailment, at each time increment in the event. In all applications reviewed and all methods tested, the model is fit separately for each account.

In most cases, the model is also fit separately for each event, because different data are selected. The exception is model fitting based on a full season of data. In these cases, the same model fit applies to all events. The estimated loads vary by event because the control-day conditions vary.

In these models, each observation corresponds to a particular day and hour (or finer time interval). The dependent variable is the account's load at that day and hour. In almost all the applications reviewed, a different set of coefficients is estimated for each hour of the day. The predictor variables are typically weather variables and possibly daytype. Thus, each observation consists of the account's load for a particular day and hour together with the corresponding weather variables.

The purpose of the regression model is to provide a reliable estimate of what the load for a particular customer would have been on a particular day in the absence of a control event. This is somewhat different from contexts where a model must yield reasonable results across a population of customers and over a broader time span. Inaccuracies for an individual customer and event have financial consequences for the customer, and in the long run may affect the customer's willingness to participate in a demand response program.

As noted in Section 2, factors affecting a customer's load on a particular day include the daytype, weather, and activity level. For some customers, activity level may vary from day to day in ways that have substantial effects on load. However, meaningful, objective customer-specific variables that track activity by day are typically not available. Daytype is readily available. However, we assume that attention is restricted to business days within a season. Daytype was therefore fixed within the models tested. Thus, the models tested include only weather terms.

Differences among regression models for this application include the following.

### *The Type of Weather Variable(s) Included*

Typically, temperature, degree-days, and/or temperature-humidity index are used. Some load research analysts incorporate additional weather terms such as precipitation, cloud cover, sunshine, and wind speed in load models, but these were not used in the primary demand-response estimation methods reviewed. Temperature and humidity are the dominant drivers of cooling loads.

If the facility is consistently in cooling (or heating) mode across the span of the data used in the regression, degree-day variables offer no advantage over temperature variables. However, if the data include milder conditions when cooling or heating is not required, degree-day variables generally perform better. In effect, degree-days "count" temperature only when it is high enough to require cooling or low enough to require heating.

### *Whether the Degree-Day Base is Fixed in Advance*

If degree-day variables are used, the degree-day base may be fixed in advance or may be determined from the regression. The base or reference temperature for a building is the temperature at which cooling or heating is first required in the building. Cooling or heating degree-days for a day are the difference between this reference temperature and the outside air temperature.

Degree-days calculated at the appropriate temperature are more closely correlated with load than degree-days at other bases. If too low a reference temperature is used for cooling, the model will tend to under-estimate load in hot weather, and overstate load in cooler weather. If too high a reference temperature is used, the opposite occurs.

The appropriate degree-day base varies substantially across buildings. The reference temperature depends on building thermal mass, solar gain, and internal loads. The only meaningful way to determine the best base temperature for a given building is by analysis of load data in relation to temperature data. Models that allow the degree-day base to vary tend to have lower systematic error, but also are more complex, more time consuming to fit, and less well determined if data are limited.

### *Whether Lagged Weather Terms are Included*

Lag temperature or degree-day terms are used to account for heat build-up over time in a building. This is the effect of "thermal mass." An approach the authors have used includes as a predictor the weighted average of degree-days for the past 48 hours, with the weights exponentially decreasing. Simpler approaches put multiple temperature or degree-day variables into the model, at different lags. Lagging humidity is not highly meaningful, as buildings do not store humidity in the way that they store heat.

### *Whether the Predictor Variables are Hourly or Daily*

Although the coefficients are almost always allowed to vary by hour of the day, the predictors may vary only daily. Because buildings store heat, the cooling load in a building does not respond instantaneously to the outside temperature, but depends on the temperature over recent hours. Explicit incorporation of lag weather variables can account for this effect. However, modeling load in a given hour as a function of the average temperature over the day can often work as well.

In part, this approach works because the variation in temperature over the course of the day is similar from day to day. Thus, the 10 am coefficients of daily average temperature tend to be smaller than the 4 pm coefficients. This is partly because there has been less heat build-up by 10 am and partly because the actual outdoor temperature at 10 am for a given daily average is lower than that at 4 pm.

### Models Tested

All the models we tested had hourly varying coefficients including an hourly intercept. That is, 24 sets of hourly coefficients were fit. If no terms but the intercept were included in the model, the model reduced to the average, by hour of the day, of the loads on the days included in the model fit.

The following models were tested.

A. Average. No variables besides the intercept term.
   $L_{dh} = \alpha_h$

B. Daily temperature.
   $L_{dh} = \alpha_h + \beta_h T_d$

C. Hourly temperature
   $L_{dh} = \alpha_h + \beta_h T_{dh}$

D. Daily heating and cooling degree-days
   $L_{dh} = \alpha_h + \beta_h HDD_d + \gamma_h CDD_d$

E. Hourly heating and cooling degree-hours
   $L_{dh} = \alpha_h + \beta_h HDH_{dh} + \gamma_h CDH_{dh}$

F. Hourly heating and cooling degree-hours with lagged degree-hours
   $L_{dh} = \alpha_h + \beta_{1h} HDH_{dh} + \gamma_{1h} CDH_{dh} + \beta_{2h} LHDH_{dh} + \gamma_{2h} LHDH_{dh}$

G. Hourly temperature-humidity index
   $L_{dh} = \alpha_h + \beta_h THI_{dh}$

In the formulas above, for a given account,

$L_{dh}$ = load at hour $h$ on day $d$

$T_d$ = daily average temperature (average of daily minimum and maximum) on day $d$

$HDD_d$ = heating degree-days base 65°F on day $d$

$CDD_d$ = cooling degree-days base 65°F on day $d$

$HDH_{dh}$ = heating degree-hours base 65°F at hour $h$ on day $d$

$CDH_{dh}$ = cooling degree-hours base 65°F at hour $h$ on day $d$

$LHDH_{dh}$ = lagged heating degree-hours base 65°F at hour $h$ on day $d$

$LCDH_{dh}$ = lagged cooling degree-hours base 65°F at hour $h$ on day $d$

$THI$ is the temperature-humidity index for hour $h$ on day $d$.

$\alpha_h, \beta_h, \gamma_h$ are coefficients determined by the regression, $h = 1, 2, \ldots, 24$.

The lag degree-day terms are based on lagged temperature calculated as

$$LT_{dh} = \sum_{k=1}^{48} T_{d,h-k} e^{-k/48} \Big/ \sum_{k=1}^{48} e^{-k/48} ,$$

where $LT_{dh}$ is lagged hourly temperature and $T_{dh}$ is hourly temperature and $e$ is the natural base.

The temperature-humidity index is calculated by PJM's method as:

$$THI_{dh} = T_{dh} - .55(1-RH_{dh}/100)(T_{dh}-58) \quad T_{dh} > 58 ,$$

where $RH$ is relative humidity. $THI$ equals hourly temperature below 58°F.

### *Conditional Weather Models*

Each of the weather models B through G was fit as a "conditional weather model." That is, the weather terms were kept in the model only if certain model diagnostics indicated that these terms were meaningful and well determined.

The full set of cooling heating coefficients were either all retained or all dropped from the model. Likewise, the full set of hourly heating coefficients were either all retained or all dropped from the model.

The requirements for retaining the cooling (or heating) terms in the model were the following:

1. Although some negative coefficients were allowed, the full set of cooling (or heating) coefficients must have a positive sum. That is, the overall heating or cooling effect identified in the model must at least be physically meaningful.

2. The F-statistic for including the set of cooling (or heating) coefficients must be significant at the 0.10 significance level. That is, the data must indicate that there is a consistent and reasonably well determined relationship between load and the weather terms.

### *4.1.3 Adjustment Method*

Adjustment methods tested are additive or scalar, as described in Section 2. Common adjustments are to some combination of the hours between 1 and 4 hours prior to curtailment. Adjustments tested were:

- additive to load at hours h0-1 and h0-2
- scalar to load at hours h0-1 and h0-2
- additive to load at hours h0-3 and h0-4
- scalar to load at hours h0-3 and h0-4
- weather-based adjustment of all loads based on the difference or ratio of regression-estimated load using curtailment period and baseline period THI (in scalar form, the PJM approach). PJM requires a minimum of one month's data or a full previous season. Because of limited previous season data the THI adjustment for this analysis is based on the present season of data. This adjustment is tested only with averages, not with weather models.

In the descriptions above, h0 indicates the hour beginning the curtailment period.

## 4.2 DAYS AND HOURS TO BE EXAMINED

Each account for which data were provided was classified as either curtailed or uncurtailed. Curtailed accounts were those that were participants in a curtailment or demand-response program. Uncurtailed accounts were not program participants. Some curtailed accounts had data provided for years in which no curtailment occurred.

The time periods for which measure performance was tested depended on the data set, and whether the load data were for curtailed or uncurtailed accounts.

### *4.2.1 Uncurtailed Accounts*

For uncurtailed accounts, certain days were set aside as test days. The baseline calculation methods were then used, with the non-test days' data as input, to calculate the load on the test days. Differences between actual and estimated load on the test days indicate the accuracy of the estimation method.

Selection of test days was as follows:

1. For data sets that included both curtailed and uncurtailed accounts on actual control days, the days when the curtailed accounts were curtailed defined the test days for uncurtailed accounts.
2. For data sets that included uncurtailed customers only, test days were selected based on extreme temperature.
3. For data sets that included some years with no curtailment for customers in a curtailment program, the test days for years with no curtailment were determined in the same way as for data sets with only uncurtailed accounts.

For each of the test days, method performance was tested for each of two 4-hour periods, one from 7 AM to 11 AM, the other from 11 AM to 3 PM.

### *4.2.2 Curtailed Accounts*

For curtailed accounts (i.e., accounts in curtailment or demand-response programs) in years when curtailments occurred, the performance was tested during the hours of actual curtailment periods. Curtailment periods were determined from the program records of when curtailment was called, not based on whether an individual customer curtailed or not.

There is no correct baseline against which to assess method performance for these accounts on these days. The preferred method(s) was selected based on the results of the tests on uncurtailed accounts. Other methods were assessed in comparison with the preferred method.

## 4.3 DATA USED FOR TESTING

Interval load data were provided for this study from several parts of the U.S., for both curtailed and uncurtailed accounts. A total of 646 accounts were used in the analysis. For some accounts, multiple years of data were used. The total number of accounts used in the study by region, year, and curtailment category is shown in the table below.

**Table 4-1**
**Interval Load Data Used in the Study**
**by Region, Year, and Curtailment**

| Region | Year | Total Number of Accounts | Non-Interruptible Accounts | Interruptible Accounts |
|---|---|---|---|---|
| California | 2000 | 3 | 2 | 1 |
| California | 2001 | 1 | | 1 |
| Mid Atlantic | 2001 | 33 | 17 | 16 |
| Midwest | 1998 | 28 | | 28 |
| Midwest | 1999 | 37 | | 37 |
| Midwest | 2000 | 39 | | 39 |
| Midwest | 2001 | 61 | 11 | 50 |
| Northeast | 2001 | 15 | 5 | 10 |
| Northwest | 2000 | 16 | 10 | 6 |
| Northwest | 2001 | 16 | 10 | 6 |
| Southeast | 1992 | 66 | 24 | |
| Southeast | 1998 | 66 | | 66 |
| Southeast | 1999 | 62 | | 62 |
| Southeast | 2000 | 63 | | 63 |
| Southwest | 1997 | 24 | 17 | 7 |

All the regions had accounts with summer curtailment data. Only the Midwest, Northwest, and Southeast had nonsummer curtailment data.

## 4.4 LOAD TYPE CLASSIFICATION

The results in the next section are displayed separately by load type. Load types are weather-sensitive and non-weather-sensitive, each with low variability and high variability.

Initially, we examined results separately by location. The key difference across locations appeared to be extreme errors that occurred for certain data sets that had very large accounts with wide swings in load. For this reason, we developed the load type classifications.

### *Weather Sensitivity*

Accounts were classified as weather-sensitive or not based on the diagnostics from a weather model fit. The weather model used was an hourly degree-hour model based on a full year of load data, with the degree-day bases estimated as part of the model. This is model "E" as defined above, except that the degree-day bases are parameters estimated by the model.

This "classification" model was used only for this purpose, and was not used as one of the tested methods. Because of the optimized selection of degree-day base, this model requires extensive processing time when applied to large numbers of accounts. In addition, its performance can be unreliable if restricted to short time periods.

The classification model diagnostics determined if the heating and/or cooling coefficients should be dropped from the model. This determination was made in the same way as described above for all the conditional weather models. As noted, these diagnostics required

1. The full set of cooling (or heating) coefficients must have a positive sum.
2. The F-statistic for including the set of cooling (or heating) coefficients must be significant at the 0.10 significance level.

If cooling coefficients were retained the account was considered weather sensitive for the summer analysis. If heating coefficients were retained the account was considered weather sensitive for the nonsummer analysis.

For all the models tested, these same diagnostic criteria were used for each model fit. Thus, an account classified as non-weather-sensitive based on the classification model might have heating and/or cooling terms included for one or more of the tested models. Likewise, an account classified as weather-sensitive by the classification model might have neither set of terms for a particular tested model.

An alternative way of classifying an account as weather sensitive or not would be in terms of the fraction of its maximum load or total energy use is for cooling and/or heating. That type of information is not typically available for an account. The weather-related fraction could be estimated from the fitted classification model. However, that calculation would depend on having a reliable estimate of the cooling and heating coefficients.

The approach taken here is to classify an account as weather-sensitive if there is a positive relationship between load and the weather drivers that is discernible above the other variations in load level. Thus, an account that in fact has a large cooling load but for which other variations swamp the variations due to weather would be classified as non-weather-sensitive for the summer analysis.

### *Load Variability*

Accounts were also classified as high or low variability. Variability was assessed not in terms of how "flat" the load was across the day, but how much the load at a given hour varied from day to day. For loads that are more highly variable in this sense, any projection based on previous days is likely to have greater error. That is, baselines and corresponding demand reduction estimates for these accounts will be subject to greater uncertainty.

The account variability was measured in terms of the root-mean-square deviation of load in each hour from the corresponding mean for that hour. This root-mean-square deviation was calculated across all curtailment hours, and normalized by dividing by the root-mean-square load during these hours. This is a version of Theil's U statistic, described in Section 4.5.1, applied to the load level during curtailment periods. This statistic is similar to a coefficient of variation for load during peak hours. Because we are looking at deviations relative to the mean for each hour of the day, systematic differences across hours in the day do not affect this measure of variability, but differences in load from day to day do.

Within each season and curtailment type (summer or nonsummer, curtailed or noncurtailed) the cut-off between high and low variability was set so that approximately one quarter of the accounts were in the high variability group. The cut-offs are indicated in the table below.

**Table 4-2**
**Cut-Offs of Theil's U for High Variability Accounts**

| | Cutoff |
|---|---|
| Summer Uncurtailed | 0.29 |
| Summer Curtailed | 0.40 |
| Nonsummer Uncurtailed | 0.32 |
| Nonsummer Curtailed | 0.42 |

The cut-offs used were in the neighborhood of 30 to 40 percent for each group. For summer curtailed accounts, for example, high variability accounts are those for which the day-to-day variation from the average for each hour is greater than 40 percent. Thus, if the average weekday 4 pm load is 600 kW, the typical day-to-day variation in the 4 pm load would be more than 240 kW.

### *Data-Base Classification Compared with Account Identifiers*

We asked the data donors to provide an indication if possible as to whether each account was weather-sensitive or not. Some donors indicated that all accounts were industrial, therefore non-weather-sensitive. Others provided SIC or NAICs codes for each account. Some provided no information.

The table below shows the proportion of industrial and non-industrial accounts studied for each of the load types. While there is a general perception that industrial loads are non-weather-sensitive and commercial loads such as office are weather-sensitive, the classification based on the data put many industrial accounts in the weather-sensitive category. This finding is consistent with observations in previous work by members of the study team. However, the table also shows that the curtailed accounts were more likely to be classified as non-weather-sensitive than the uncurtailed accounts.

**Table 4-3**
**Load Type Compared with Industrial/Nonindustrial Classification**

| **Analysis Load Type** | **Industrial/Nonindustrial Classification** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Load Type | Summer Noncurtailed | | | Summer Curtailed | | Nonsummer Noncurtailed | | | Nonsummer Curtailed | |
| | Non-Ind. | Ind. | Unknown | Non-Ind. | Ind. | Non-Ind. | Ind. | Unknown | Non-Ind. | Ind. |
| Weather Sensitive Low Variability | 28 | 25 | 77 | 19 | 29 | 12 | 12 | 46 | 3 | 11 |
| Weather Sensitive High Variability | 4 | 4 | 13 | | 8 | 3 | 2 | 7 | | 5 |
| Non-Weather Sensitive Low Variability | 14 | 20 | 17 | 4 | 41 | 22 | 23 | 54 | 2 | 43 |
| Non-Weather Sensitive High Variability | 1 | 11 | 14 | 1 | 25 | | 17 | 14 | 1 | 14 |

The next tables show the numbers of accounts in each category, and their size distribution. The distributions are all quite skewed, with the mean kW greater than the 75$^{th}$ percentile (3$^{rd}$ quartile) in almost all cases. Though we attempted to get similar size accounts in the curtailed and uncurtailed groups, the tables show that the uncurtailed accounts tended to be larger.

**Table 4-4**
**Summer Uncurtailed Account Size Distribution (kW) by Load Type Classification**

| Load Type | Number of Accts | Mean | 1st Quartile | Median | 3rd Quartile |
|---|---|---|---|---|---|
| Weather Sensitive Low Variability | 130 | 1,839 | 285 | 539 | 1,354 |
| Weather Sensitive High Variability | 21 | 1,132 | 223 | 412 | 782 |
| Non-Weather Sensitive Low Variability | 51 | 4,679 | 581 | 1,337 | 3,948 |
| Non-Weather Sensitive High Variability | 26 | 4,232 | 426 | 591 | 1,408 |

**Table 4-5**
**Summer Curtailed Account Size Distribution (kW) by Load Type Classification**

| Load Type | Number of Accts | Mean | 1st Quartile | Median | 3rd Quartile |
|---|---|---|---|---|---|
| Weather Sensitive Low Variability | 48 | 5,374 | 844 | 1,868 | 4,072 |
| Weather Sensitive High Variability | 8 | 8,064 | 582 | 1,229 | 5,781 |
| Non-Weather Sensitive Low Variability | 45 | 12,607 | 1,880 | 4,915 | 9,922 |
| Non-Weather Sensitive High Variability | 26 | 10,689 | 1,142 | 2,819 | 5,933 |

**Table 4-6**
**Winter Uncurtailed Account Size Distribution (kW) by Load Type Classification**

| Load Type | Number of Accts | Mean | 1st Quartile | Median | 3rd Quartile |
|---|---|---|---|---|---|
| Weather Sensitive Low Variability | 70 | 1,313 | 263 | 495 | 1,060 |
| Weather Sensitive High Variability | 12 | 550 | 103 | 276 | 639 |
| Non-Weather Sensitive Low Variability | 99 | 3,844 | 504 | 986 | 2,462 |
| Non-Weather Sensitive High Variability | 31 | 3,603 | 412 | 1,158 | 4,166 |

**Table 4-7**
**Winter Curtailed Account Size Distribution (kW) by Load Type Classification**

| Load Type | Number of Accounts | Mean | 1st Quartile | Median | 3rd Quartile |
|---|---|---|---|---|---|
| Weather Sensitive Low Variability | 14 | 11,710 | 1,354 | 2,381 | 8,849 |
| Weather Sensitive High Variability | 5 | 2,991 | 1,142 | 1,416 | 1,561 |
| Non-Weather Sensitive Low Variability | 45 | 11,626 | 1,625 | 3,852 | 8,623 |
| Non-Weather Sensitive High Variability | 15 | 11,268 | 1,313 | 3,870 | 6,967 |

The next table shows the numbers of accounts, account-day combinations, and account-hour combinations for which baseline errors were calculated, by season, curtailment, and load type. For each account, only actual curtailment days and hours, or those defined as test periods for uncurtailed accounts, are included in the account-day and account-hour totals.

**Table 4-8**
**Number of Observations in the Study**

| | Weather Sensitivity | Variability | Accounts | Account-Days | Account-Day Hours |
|---|---|---|---|---|---|
| **Summer Uncurtailed Accounts** | WS | Low | 130 | 334 | 1,336 |
| | NWS | Low | 51 | 121 | 484 |
| | WS | High | 21 | 59 | 236 |
| | NWS | High | 26 | 47 | 188 |
| | **TOTAL** | | **228** | **561** | **2,244** |
| **Summer Curtailed Accounts** | WS | Low | 48 | 292 | 1,591 |
| | NWS | Low | 45 | 365 | 2,410 |
| | WS | High | 8 | 49 | 227 |
| | NWS | High | 26 | 192 | 973 |
| | **TOTAL** | | **127** | **898** | **5,201** |
| **Nonsummer Uncurtailed Accounts** | WS | Low | 70 | 166 | 664 |
| | NWS | Low | 99 | 263 | 1,052 |
| | WS | High | 12 | 18 | 72 |
| | NWS | High | 31 | 76 | 304 |
| | **TOTAL** | | **212** | **523** | **2,092** |
| **Nonsummer Curtailed Accounts** | WS | Low | 14 | 96 | 890 |
| | NWS | Low | 45 | 320 | 2,356 |
| | WS | High | 5 | 25 | 106 |
| | NWS | High | 15 | 120 | 459 |
| | **TOTAL** | | **79** | **561** | **3,811** |
| **All Accounts** | | | **646** | **2,543** | **13,348** |

WS=Weather Sensitive, NWS=Non-Weather Sensitive
Account/Day/Hours for Uncurtailed Accounts represent only one of the two simulated four hour control periods.

The number of accounts is the number of distinct customers for which the methods were tested. To a certain extent, method performance will be a characteristic of the customer, related to the customer's operating practices and associated load patterns. The number of account-days is the number of distinct combinations of customers and control events. Finally, the number of account-hours is the number of individual observations of method error in the analysis.

## 4.5 PERFORMANCE MEASURES

The tests were run for several alternative methods on several different data sets for several hours on several days. Developing meaningful measures of method performance is essential to provide a basis for conclusions. Performance measures provided are somewhat different for uncurtailed and curtailed loads.

### *4.5.1 Performance Measures for Uncurtailed Test Periods*

#### *Key Calculations*

For each account used as an uncurtailed test case, the estimated load for each hour and several measures of error were calculated and reviewed for each method. Hours included in the calculations were actual curtailment hours for curtailed accounts, and test period or “simulated curtailment” hours for uncurtailed accounts. The key calculations used and displayed in this report are shown in Table 4-9.

**Table 4-9**
**Key Calculations for Each Estimation Method and Account**

| Description | Formula |
|---|---|
| estimated load | $\hat{L}_{jdh}$ = estimated load for account j on day d at hour h |
| hourly error | $e_{jdh} = L_{jdh} - \hat{L}_{jdh}$ |
| relative hourly error | $r_{jdh} = e_{jdh}/L_{jdh}$ |
| account root-mean-square hourly error | $RMSEH_j = \sqrt{[\Sigma_{d=1}^{nj} \Sigma_{h=1}^{njd} e_{jdh}^2 / \Sigma_{d=1}^{nj} n_{jd}]}$ |
| account root-mean-square hourly load | $RMSLH_j = \sqrt{[\Sigma_{d=1}^{nj} \Sigma_{h=1}^{njd} L_{jdh}^2 / \Sigma_{d=1}^{nj} n_{jd}]}$ |
| account Theil’s U hourly, or relative root-mean-square hourly error | $TUH_j = RMSEH_j / RMSLH_j$ |
| Theil’s U daily, or relative root-mean-square daily error | $TUD_j = RMSED_j / RMSLD_j$ |
| account average error | $e_{j..} = \Sigma_{d=1}^{nj} \Sigma_{h=1}^{njd} e_{jdh} / \Sigma_{d=1}^{nj} n_{jd}$ |
| relative account error | $r_{j\_\_} = e_{j..}/L_{j..}$ |

In the equations above,

$L_{jdh}$ = actual load for account *j* on day *d* at hour *h*.
$n_{jd}$ = number of hours in curtailment period for account *j* on day *d*
$n_j$ = number of test days for account *j*.

### *Key Indicators of Method Accuracy*

The goal of the performance tests is to assess the accuracy of the various baseline measures tested. As discussed in Section 3.3, accuracy has two components, bias and variability. Bias is the systematic tendency to over- or under-state baselines and corresponding load reductions. Variability is how widely the baseline is likely to vary from its typical or expected level.

#### Bias

As a key measure of bias, we focus on the median relative hourly error. Because accounts in this study are of widely varying sizes, when looking at the range of errors across accounts we need to normalize them. Thus, the error for each hour for each account is expressed as a fraction of the actual load for that hour and account. If the median, across all accounts and curtailment hours, of these relative hourly errors is positive, then more often than not the baseline is overstated, and the magnitude of demand response is overstated. If the median of the relative hourly errors is negative, then more often than not the baseline is understated and the magnitude of demand response is understated.

**Overall Error Magnitude**

As a key measure of the total magnitude of error, we consider Theil's U statistic for each account. This statistic is a "relative root-mean-square error." It is calculated as the ratio of the root-mean-square error to the root-mean-square load.

The root-mean-square error is like a standard deviation, and represents the typical error magnitude for the account. This root-mean-square error reflects both systematic error, or bias, and the level of variability around the typical error.

The root-mean-square load is a corresponding "typical" load level. Normalizing the root-mean-square error by the root-mean-square load is something like calculating a correlation coefficient. However, the U statistic may be greater than 1, since errors can be greater than the loads they estimate.

The denominator of the U statistic will be weighted toward larger loads. That is, for a given account, relative errors at higher-load hours will tend to count more heavily than those at lower-load hours.

Theil's U statistic calculated for a given account indicates the typical relative error magnitude for that account. The distribution of this statistic across accounts indicates the range of performance. We look at this distribution in terms of both the median and an extreme, the 95$^{th}$ percentile. The median Theil's U indicates the typical relative error magnitude for a typical account. The 95$^{th}$ percentile indicates performance in the worse cases.

Descriptive statistics and graphical displays of distributions of the error measures are presented in Section 5.

### 4.5.2 Performance Measures for Curtailment Days

For the curtailed load during curtailment periods, an approximate error $e^{\sim}_{jdh}$ was calculated for each method as the difference between the load estimate from that method and the preferred method. That is,

$$e^{\sim}_{jdh} = L^{\wedge}_{jdh} - L^{*}_{jdh} \ ,$$

where

$L^{*}_{jdh}$ = estimated load for account *j* on day *d* and hour *h* using the preferred method.

This calculation was performed only for day-hour combinations *dh* included in called curtailment periods for account *j*. All of the error measures used for the uncurtailed tests were then calculated for the curtailment periods.

In addition, for the curtailment periods the load reduction was calculated for the preferred method and for each other tested method:

$$D_{jdh} = L^{*}_{jdh} - L_{jdh}$$

$$\hat{D}_{jdh} = \hat{L}_{jdh} - L_{jdh}\,.$$

The error in the demand response is the same as the error in the load estimate:

$$\hat{D}_{jdh} - D_{jdh} = (\hat{L}_{jdh} - L_{jdh}) - (L^{*}_{jdh} - L_{jdh}) = \hat{L}_{jdh} - L^{*}_{jdh} = \tilde{e}_{jdh}\,.$$

However, relative errors are of interest not just relative to the load level but also relative to the demand response $D$. Two types of relative errors were calculated for the curtailed accounts. The first is the same as for uncurtailed accounts, expressing the hourly error $\tilde{e}_{jdh}$ as a fraction of the hypothetical load in the absence of curtailment (based on the preferred method). The second type of relative error is the hourly error as a fraction of the curtailment amount determined using the preferred method.

# 5 RESULTS OF PERFORMANCE TESTS

## 5.1 INTRODUCTION

This section presents the results of the tests described in Section 4. The purpose of the tests is to examine the accuracy of baselines determined by various methods, under a wide range of conditions. The baseline methods tested and the data sets on which they were tested are described in Section 4.

For each method tested, the key measures we use to examine the method performance are the following.

**Bias: Median relative hourly error**, across all accounts, test days, and hours**.**
The relative hourly error $r_{jdh}$ is calculated for each account-day-hour combination. The median of all these relative hourly errors is the key indicator of method bias. Median relative hourly error greater than 0 indicates a systematic tendency to overstate baselines and load reductions. Median relative hourly error less than 0 indicates a systematic tendency to understate baselines and load reductions.

**Typical error magnitude: Median U statistic**, across all accounts.
This measure reflects both bias and variability. For each method, for each account, Theil's U is calculated across all test days and hours. The U statistic is similar to a coefficient of variation. It indicates the typical error magnitude relative to a typical load level. The median of U across accounts indicates the typical relative error for a typical account.

**Error magnitude for extreme accounts: 95$^{th}$ percentile U statistic,** across all accounts.
The extremes of Theil's U indicate how far off the baseline tends to be for the less well-behaved accounts. The focus is not on individual problematic hours, but on accounts for which the method error tends to be high across hours.

For curtailed accounts, the same types of error measures are calculated as for noncurtailed accounts, except that each of these is calculated for errors relative to estimated load and also for errors relative to estimated curtailment amounts.

## 5.2 A FIRST LOOK AT THE COMPARISONS

Before considering these performance measures across the full range of methods explored, we illustrate the types of comparisons being made. For purposes of this illustration, we consider two combinations of estimation method and data selection rule:

- A1 is the average, by hour of the day, of the past 10 uncurtailed business days.

- D8 is a regression model using a full season of data, and daily heating and/or cooling degree-days as the predictor variables for each hour. Like all the weather models tested, it is a “conditional” model, with heating or cooling degree-days included in the model for a particular account only if the diagnostic screens are passed.

Formal specifications of these methods are given in Section 4.

For each of these combinations of estimation method and data selection criteria, each of the possible adjustments was applied.

- Additive adjustment to hours h0–1 to h0–2
- Scalar adjustment to hours h0–1 to h0–2
- Additive adjustment to hours h0–3 to h0–4
- Scalar adjustment to hours h0–3 to h0–4
- Additive adjustment based on the THI regression model using the previous season’s data.
- Scalar adjustment based on the THI regression model using the previous season’s data.

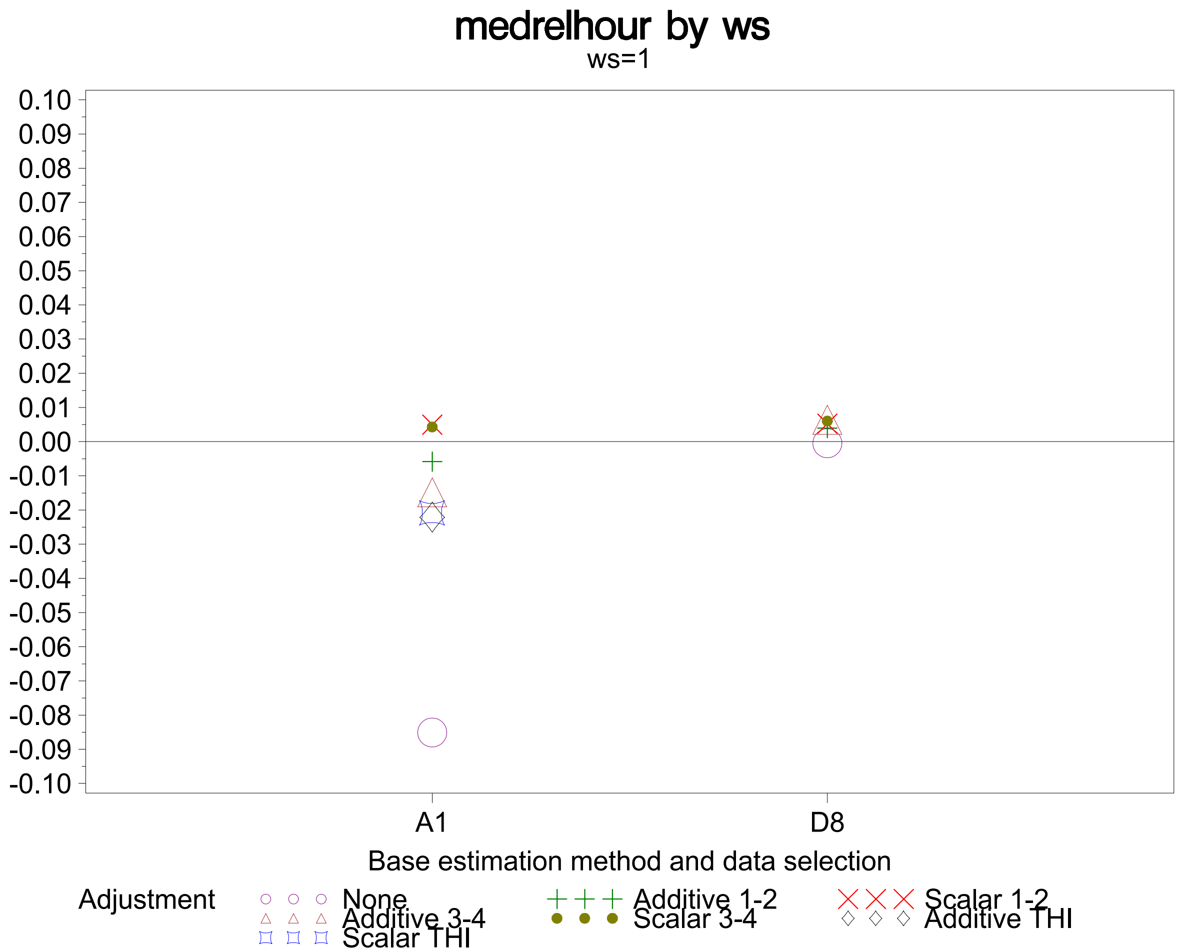
These adjustments are described and illustrated in Section 2.

### 5.2.1 Looking at Bias

We focus on the median relative hourly error as the key measure of bias, that is, systematic tendency to over- or understate load. The median rather than mean relative hourly error is used because relative errors can be quite large if the true load is small, and these occasional explosive values can distort the mean without indicating typical behavior.

Figure 5-1 shows the median relative hourly error for the average A1 together with each of the adjustments. The results for the different adjustments are plotted in a vertical line, and indicated with different symbols. The figure shows that the unadjusted method A1 (circle) has a median relative hourly error around –8 percent. That is, the simple average of the last 10 uncurtailed business days tends to understate the baseline by about 8 percent.

**Figure 5-1**
**Median Relative Hourly Error**
**Simple 10-Day Average and Full-Season Degree-Day Regression**
**Summer Uncurtailed Weather-Sensitive Low-Variability Accounts**




With the additive adjustment to the last two hours before curtailment ('+') the median relative error is still negative, but only about half a percent. With an additive adjustment to the third and fourth hour before (triangle) yields a relative error of –1.5 percent. Using the more complicated THI adjustment in additive (diamond) or scalar (square) gives about –2 percent. Scalar adjustment to the last two hours ('x') or to the third and fourth hour (black dot) before give slightly positive bias.

Overall, then, any of the adjustments seem to reduce substantially the negative bias of using a simple average of the past 10 days. That is, use of these adjustments mitigates the average's understatement of baselines and demand reductions for weather-sensitive accounts.

Also shown in the figure is the same set of results for method D8, the full-season daily degree-day model. The figure shows that the unadjusted weather model ('o') has only slight negative bias, -0.5 percent. With any of the adjustments, the model has a slight positive bias, about +0.5

percent. The THI adjustments were not tested with the regression models, on the assumption that the models themselves should capture whatever effect is captured in the THI adjustment. Thus, for weather-sensitive loads, use of the full-season model substantially eliminates any systematic tendency to over- or under-state the baseline.

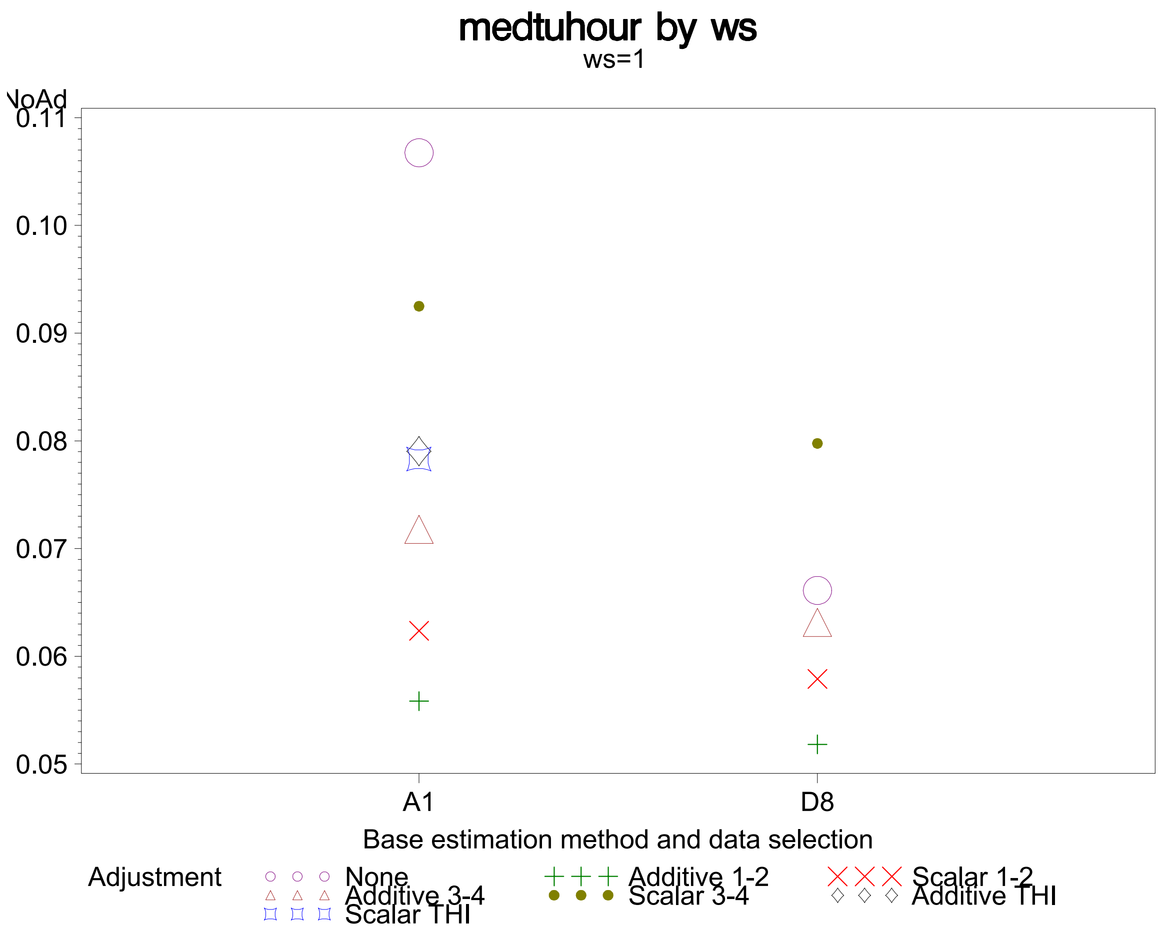
### *5.2.2 Looking at Overall Error Magnitude*

The key indicator of error magnitude we use is the median, across all accounts, of Theil's U statistic for the account. Theil's U is the root-mean-square error, relative to the root-mean-square load. Thus, high relative errors for hours when the load is small contribute little to U. Relative errors at hours when the load is high count most heavily. The median of Theil's U across accounts indicates what the typical hourly error is for a typical account. Theil's U reflects both systematic error and variability.

Figure 5-2 shows the medians of Theil's U for the same cases whose median relative errors are plotted in Figure 5-1. For the unadjusted average A1, the median account U is about 11 percent. All of the adjustment methods reduce U. The lowest U for method A1 is with the additive and scalar adjustments to hours –1 to –2, at 5.4 and 5.8 percent, respectively.

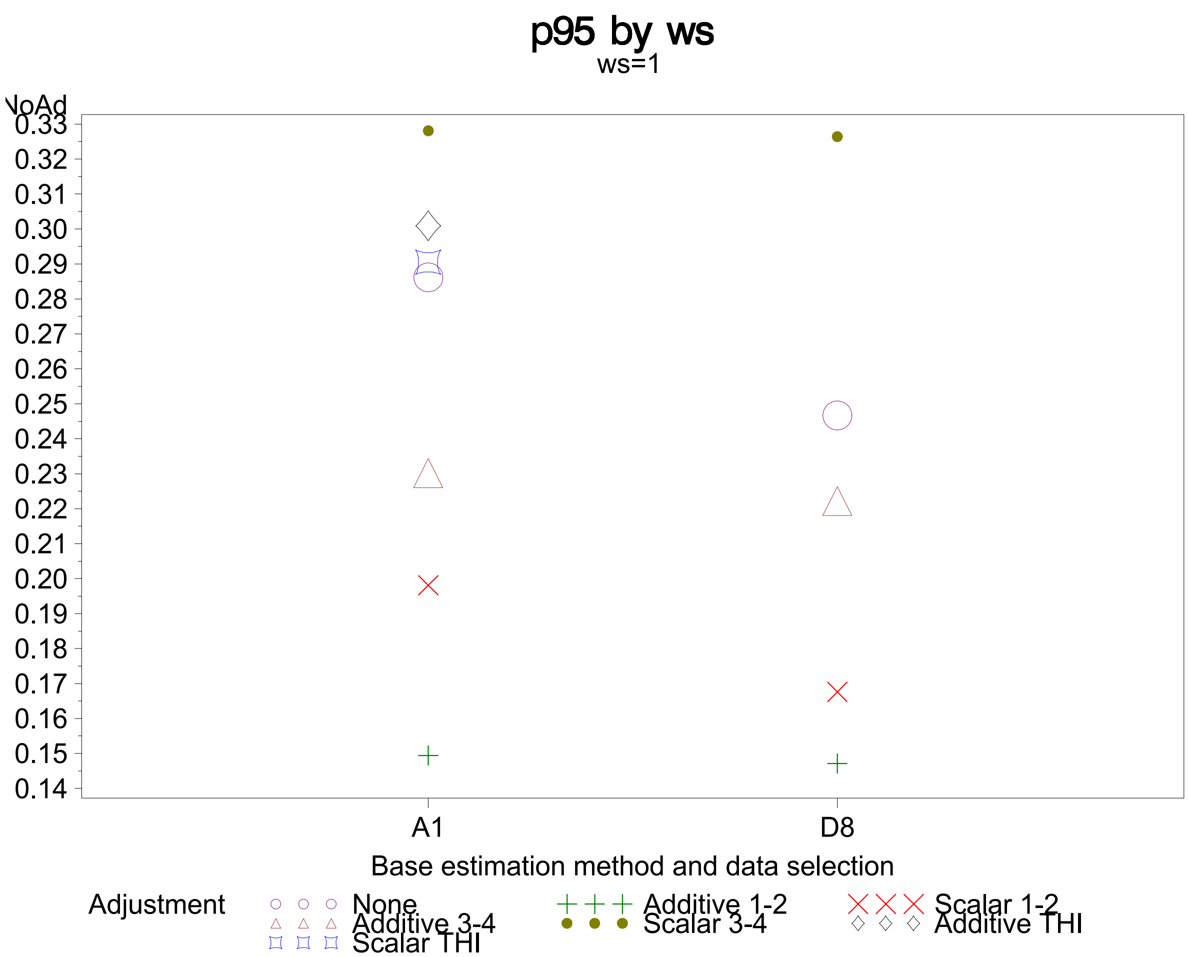
For the weather model D8, the unadjusted method has Theil's U of 6.2 percent, only slightly worse than the best adjusted average. The additive and scalar adjustments to the last two hours are slightly better, at 4.9 and 5.7 percent, respectively. The scalar adjustment to hours –3 to –4 is somewhat worse, at 7.5 percent.

**Figure 5-2**
**Median Account Theil's U**
**Simple 10-Day Average and Full-Season Degree-Day Regression**
**Summer Uncurtailed Weather-Sensitive Low-Variability Accounts**




In addition to looking at the median of Theil's U, we also looked at the 95th percentile. This measure indicates how bad the errors tend to be for the less predictable accounts. Figure 5-3 shows these results for the same cases as in Figures 5-1 and 5-2. Because we are looking at the extremes, the scale of the plot is changed to accommodate the values.

**Figure 5-3**
**95th Percentile Account Theil's U**
**Simple 10-Day Average and Full-Season Degree-Day Regression**
**Summer Uncurtailed Weather-Sensitive Low-Variability Accounts**




The figure shows that the extreme cases show typical error in the range of 15 to 35 percent, depending on the adjustment method and estimation method. The 95th percentiles of Theil's U show much more spread across adjustment methods than do the medians. Nevertheless, some of the same patterns observed for the medians still hold:

- Without adjustment, the full-season weather model performs better than the 10-day average.
- For both average and regression, the additive adjustment to the last two hours has smallest errors, and the scalar adjustment the next smallest.
- With the additive adjustment to the last two hours, the simple average is nearly as good as the weather model.

However, the difference between the better- and worse-performing adjustments is more pronounced for the extreme cases than for the medians.

### 5.2.3 Preliminary Indications

Based on these illustrative results, some preliminary indications are the following:

1. A relatively simple adjustment to a simple method may yield a method that performs nearly as well as a more complicated weather model.
2. Even for a complicated weather model, there may still be some bias if no adjustment is made.
3. The weather model appears to be less sensitive to the type of adjustment made than is the simple average.
4. The additive adjustment to the two hours before curtailment appears to perform best in terms of both bias and variability.
5. In general, the patterns in the 95$^{th}$ percentile are qualitatively similar to those for the median, though the magnitudes are greater, and the corresponding difference between the better and worse methods is greater.

As will be seen, these patterns are borne out in other comparisons. While we might not have anticipated all these results, they do make sense.

Some of the reasons that adjustment to the hours just before curtailment could improve baseline accuracy were discussed conceptually in Section 2. These hours are most reflective of actual operations just prior to curtailment. Thus, to the extent operations on different from those in the baseline days, the adjustment to the pre-curtailment hours captures some of that difference. These hours also reflect the actual weather on that day, at least up to the time of the curtailment. As a result, this adjustment can also serve to correct for weather differences, to an extent.

It is not surprising that adjustment to the two hours just before curtailment would perform somewhat better than the adjustment to the third and fourth hour before. The later hours are more likely to be past the morning ramp-up, reflect more of the weather and heat build-up for the day, and be indicative of operations for the main part of the day.

Not necessarily expected was the finding that the additive adjustment does slightly better than the scalar adjustment overall. This result appears to be due to the scalar adjustment's becoming very large if the baseline load was very low compared to the actual during the adjustment hours.

Also not necessarily expected was the finding that the THI adjustments tend to be worse in terms of both bias and overall error magnitude than the adjustment to recent hours. This result suggests there are important curtailment-day differences apart from weather, which are captured in the other adjustments. The THI adjustment has the advantage of being "ungamable," because it does not utilize recent loads. Its disadvantage is that it is therefore not informed by differences in these loads when they might be meaningful indicators of operations on the curtailment day.

Particularly intriguing is the performance of the full-season weather model with additive adjustment to the last two hours. The unadjusted model has near zero bias, but the adjustment noticeably reduces the variability of the method, bringing the error magnitude down from 6.2 percent to 4.9 percent in terms of median account U, and from 25 percent to 15 percent for the extreme (95$^{th}$ percentile) accounts.

While it might have been expected that the weather model would not benefit from the simple adjustments, this result does make sense. It is also consistent with the finding that the weather-based THI adjustments don't do as well as the adjustment to the last two hours. That is, there are non-weather effects that are difficult to model explicitly, but that are reflected in the earlier load on the day being predicted.

Another way to say this is that the prior hours capture some of the random effects that are left unexplained by the weather model. Load forecasters often include the most recent data as available as explicit terms in prediction models. Rather than including the adjustment as an extra tweak on the weather model, the load in the hours prior to curtailment could be included as predictors in the model. Model diagnostics could then determine whether to retain these terms, while the model coefficients would determine the relative weight given to the prior data and the current-day information.

## 5.3 A Brief Guide to the Displays

In the remainder of this chapter, we present the same types of figures as Figures 5-1 through 5-3. However, in each figure we include all tested combinations of estimation method, data selection criteria, and adjustment.

### *Combinations Tested*

In the displays, each combination of an estimation method and data selection criteria is indicated by a letter code for the method together with a number code for the selection criteria. Brief definitions of the codes are indicated in the table below. These method components are described more fully in Section 4.

**Table 5-1**
**Codes for Estimation Methods and Data Selection Criteria**

| Method Code | Estimation Method |
|---|---|
| A | Average |
| B | Daily Temperature |
| C | Hourly Temperature |
| D | Daily Degree-Day |
| E | Hourly Degree-Day |
| F | Lagged Hourly Degree-Day |
| G | THI |

| Selection Code | Data Selection Rule |
|---|---|
| 1 | Previous 10 |
| 2 | Previous 11 |
| 3 | Previous 10 starting d0-2 |
| 4 | Previous 20 |
| 5 | Previous 10 and Next 10 |
| 6 | High 10 of 11 |
| 7 | High 5 of 10, starting d0-2 |
| 8 | Full season |
| 9 | Full previous season |

Each estimation method and data selection criteria was tested with each of the possible adjustments listed in Section 4 and displayed in Figures 5-1 through 5-3. The exception is that the THI adjustments were applied only to averages, not to weather models.

The component methods selected were designed to include the main ISO methods in use. The codes corresponding to each of these methods are indicated in the table below. Also indicated are ways the data selection rule tested differs from the details of the ISO method.

**Table 5-2**
**Correspondence of Tested Methods to ISO Baseline Methods**

| | ISO Baseline Methodology | | | | | Closest Method Tested | | | | | | | Differences between Tested method and ISO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Data Selection | | | Estimation Method | Adjustment | Estimation Selection Code | Shorthand Descriptions | | | Full Data Selection Description | | | |
| ISO | Initial | Final | Exclusions | | | | Estimation Method | Selection Method | Adjustment | Initial | Final | Exclusions | |
| California #1 | Previous 10 business days beginning on d0-1 | Same | Control days, with replacement | Average | None | A1 | Average | Last 10 | None | Previous 10 business days beginning on d0-1 | Same | Control days, with replacement | |
| California #2 | Previous 11 business days beginning on d0-1 | Days with top 10 average loads for control period | Control days, with replacement | Average | None | A2 | Average | High 10 of 11 | None | Previous 11 business days beginning on d0-1 | Days with top 10 average loads for control period | Control days, with replacement | |
| New England | Previous 10 business days beginning on d0-1 | Same | Control days and extreme load days, without replacement (min. 7 days) | Average | Additive 1-2 | A1 | Average | Last 10 | Additive 1-2 | Previous 10 business days beginning on d0-1 | Same | Control days, with replacement | Control days are replaced with replacement (always ten days). No extreme load day exclusions. |
| New York | Previous 10 business days beginning on d0-2 | Days with top 5 average loads for control period | Control days and extreme load days, with replacement | Average | Optional Scalar 3-4 (limited to .8 to 1.2) | A7 | Average | High 5 of 10 | Scalar 3-4 | Previous 10 business days beginning on d0-2 | Days with top 5 average loads for control period | Control days, with replacement | No extreme load day exclusions. Unbounded adjustment. |
| PJM | Previous 10 business days beginning on d0-2 | Days with top 5 average loads for control period | Control days and extreme load days, with replacement | Average | Optional Scalar THI adjustment | A7 | Average | High 5 of 10 | Scalar THI | Previous 10 business days beginning on d0-2 | Days with top 5 average loads for control period | Control days, with replacement | No extreme load day exclusions. THI adjustment parameters are estimated on full same season rather than previous season or partial same season. |

### *Organization of the Displays*

Each display consists of a plot similar to one of Figures 5-1 through 5-3. Combinations of estimation method and data selection criteria are indicated by horizontal position in the graph. The adjustment method is indicated by the plotting symbol.

**Bias Plots**

Each plotted point is the median relative hourly error for a different combination of method features. From left to right, different combinations of estimation method and data selection criteria are shown. For each of these combinations, the median relative hourly error is plotted for each of the adjustment methods tested.

**Typical Error Magnitude Plots**

The median of the U statistic across accounts is plotted for each method. The different methods are displayed in the same way as for the median relative hourly error.

**Error Magnitude for Extreme Accounts**

This display is the same as the median account U, except that the 95$^{th}$ percentile is plotted instead of the median.

A different set of figures is presented for each combination of

- curtailment classification (curtailed or uncurtailed accounts)
- load type (weather-sensitive or non-weather-sensitive accounts, with high or low variability)
- season (summer or nonsummer).

Thus, a total of 16 sets of the three figures are presented. We begin with summer uncurtailed accounts. We first look at weather-sensitive accounts, low and then high variability. We then examine non-weather-sensitive accounts. This same sequence is then followed for nonsummer uncurtailed, summer curtailed, and nonsummer uncurtailed.

## 5.4 TEST RESULTS FOR UNCURTAILED ACCOUNTS, SUMMER

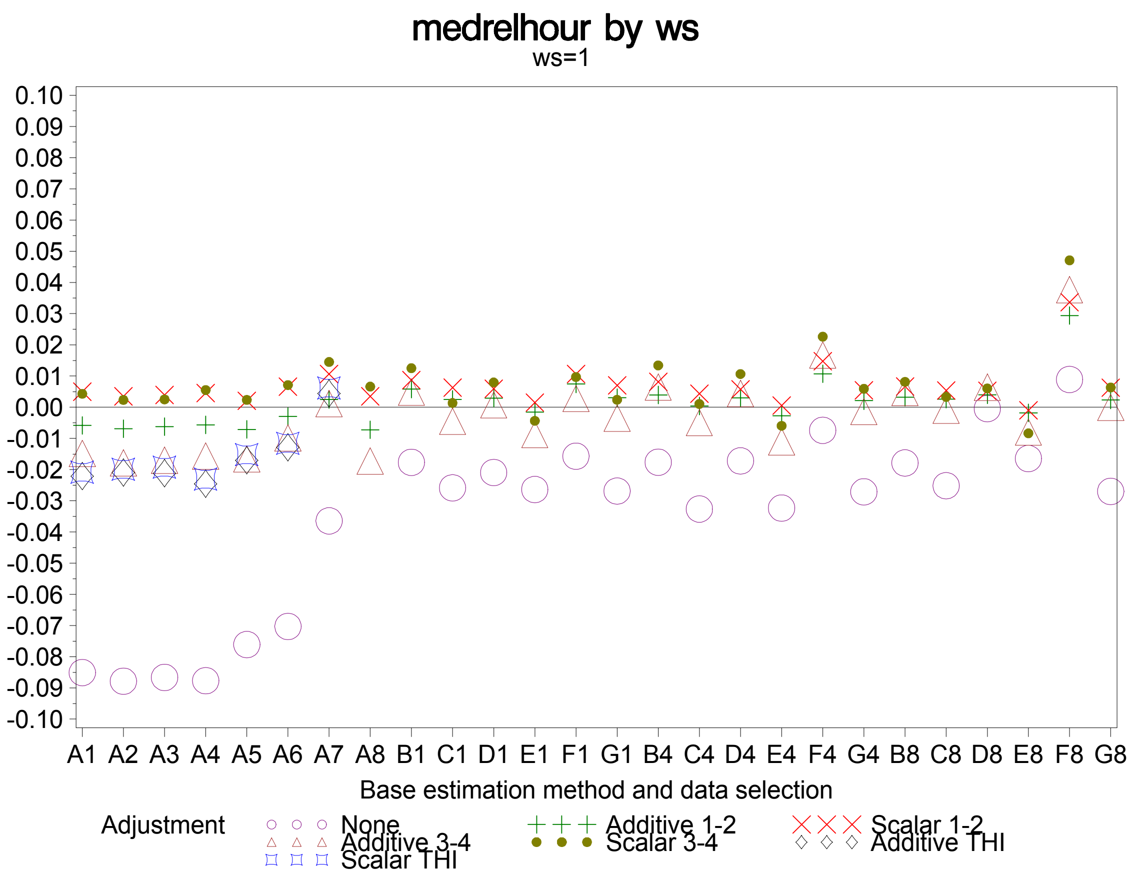
We begin with the uncurtailed accounts, summer season. For uncurtailed accounts, the actual load during the test period is known. Thus, the estimation methods can be tested against actual loads. Method performance based on these cases can be used to select the best method. For the curtailed accounts, performance of each method can then be assessed relative to "truth" as defined by this best method. We focus first on the summer because the majority of load reduction programs are aimed at reducing summer peaks.

### 5.4.1 Weather-Sensitive Loads

#### Low-Variability Accounts

Figure 5-4 shows the median relative hourly errors for all tested methods, for weather-sensitive low-variability accounts. Starting from the left, the first group of methods shown are the various averages, based on different data selection criteria. Next are the different weather models, all using the last 10 uncurtailed business days. Next is the same sequence of models, all using the last 20 uncurtailed business days. Finally, the same sequence of models use a full season of data. For each estimation method and data selection criteria, the results with the different adjustments are plotted using the same symbols as in Figures 5-1 through 5-3.

**Figure 5-4**
**Median Relative Hourly Error**
**Summer Uncurtailed Weather-Sensitive Low-Variability Accounts**




### Bias

While there are a large number of factors being compared in the figure, some patterns can be observed.

1. With no adjustment, the weather models all have lower magnitude bias than any of the averages. However, the "High 5 of 10" method (A7, highest 5 of the last 10 uncurtailed business days) is nearly as good.
2. Except for the High 5 of 10 method, the averages all have much worse bias (-6 to –10 percent) than the weather models (on the order of –3 percent or smaller).
3. With no adjustment, almost all the weather models still have some negative bias. The one exception is the full-season lag model (F8). This method becomes slightly positively biased when adjustments are incorporated.
4. With adjustment to the hours h0–1 and h0–2, all the methods have bias within ±1 percent, except the full-season lag model.
5. The adjustment to hours h0–3 to h0–4 are slightly worse than to the hours h0–1 to h0–2, and, the THI adjustments a little worse than these. The exception is the High 5 of 10

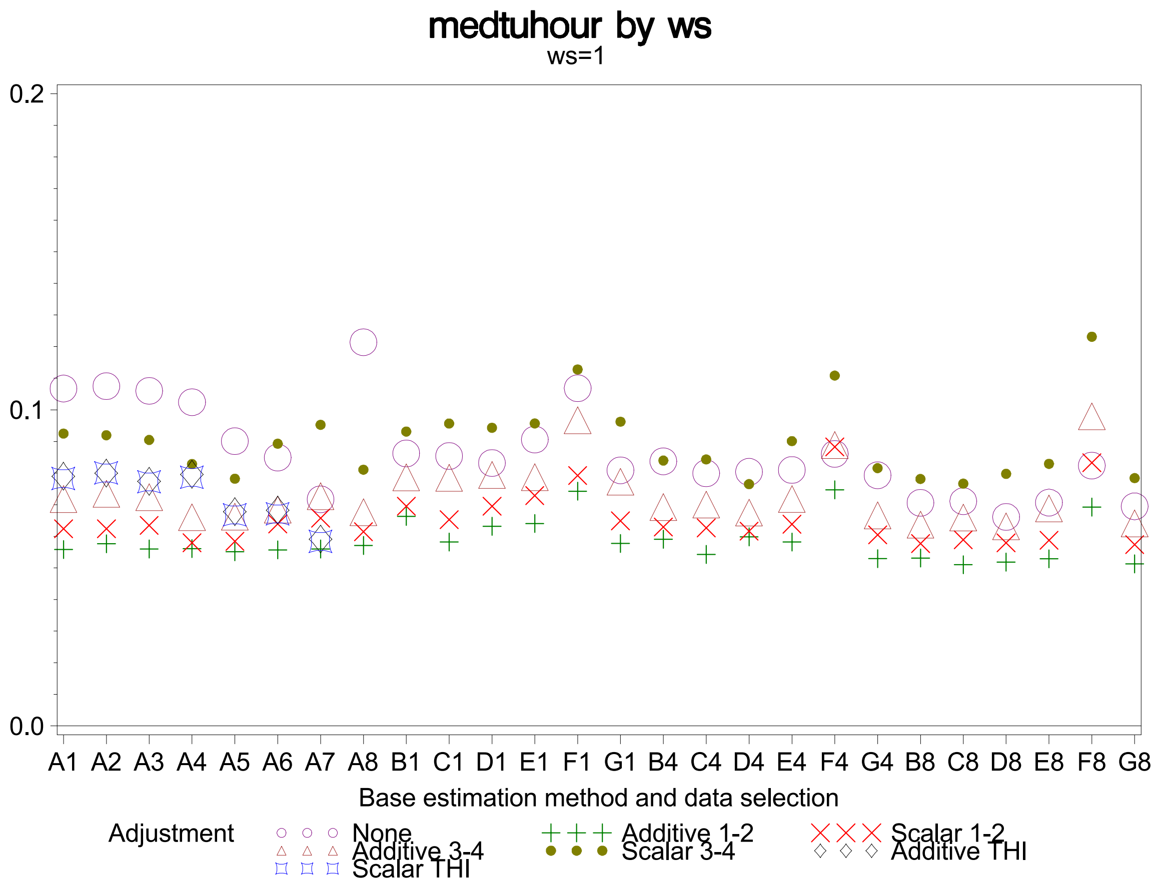
method, where the additive adjustment to hours h0–1 to h0–2 and the additive and scalar THI adjustments all give near 0 bias.

**Typical Error Magnitude**

Figure 5-5 shows the median account U for the same methods and loads. Observations from this figure are the following.

1. For each estimation-selection method, the additive adjustment to the last 2 hours has the smallest median Theil's U. That is, this combination tends to yield the smallest magnitude relative error for each account. Next best is the scalar adjustment to the last 2 hours.
2. With this additive adjustment, there is little difference among the averages A1 through A8.
3. For any adjustment and any selection rule, there is little difference among the different weather models B through G, except that the lag models F tend to have higher median Theil's U. That is, the inclusion of the lag terms in the models tends to increase the model variance.
4. With the best adjustment a1, the 20-day weather models (other than the lag F) have slightly lower error magnitude than the 10-day averages, and the full-season models still lower. However, the 10-day weather models tend to have slightly higher errors than the averages.
5. For the unadjusted methods, the averages are generally worse than the weather models. The exceptions are that the High 5 method is better than the 10- and 20-day regressions, and the lag models are as bad as some of the averages.
6. For all the weather models, the scalar adjustment to hours h0–3 to h0–4 has worse U than the unadjusted model.

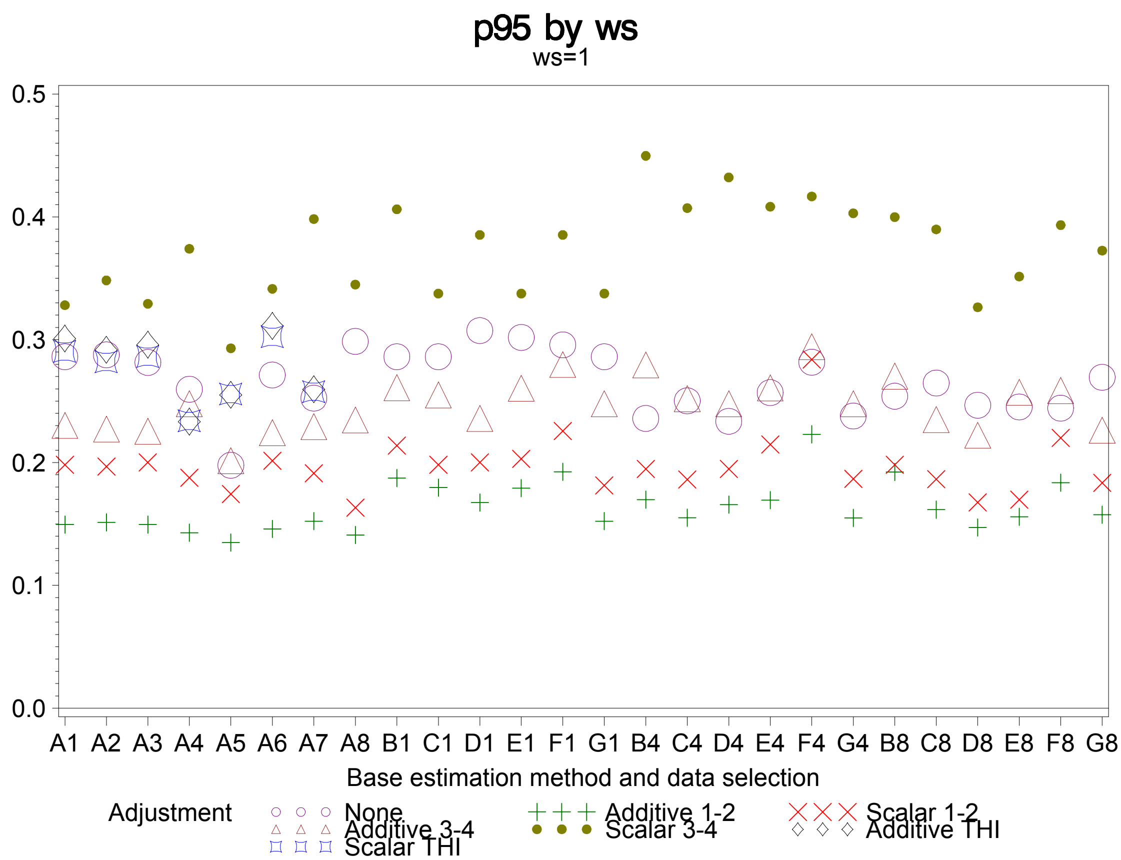
**Figure 5-5**
**Median Account Theil's U**
**Simple 10-Day Average and Full-Season Degree-Day Regression**
**Summer Uncurtailed Weather-Sensitive Low-Variability Accounts**




### Error Magnitude for Extreme Accounts

Figure 5-6 shows the 95th percentile of Theil's U for the same cases. For these extreme error magnitudes, the weather models are not dramatically better than the averages, even without adjustment. The additive adjustment to the two hours prior to curtailment performs best in almost all cases. The improvement between the unadjusted and this adjustment is substantial even for the weather models, from around 30 percent down to 15 to 20 percent typical relative error. Thus, while the adjustment choice seems to make little difference in performance for the typical account, it does make a difference to how bad the worse accounts get.

**Figure 5-6**
**95th Percentile Account Theil's U**
**Simple 10-Day Average and Full-Season Degree-Day Regression**
**Summer Uncurtailed Weather-Sensitive Low-Variability Accounts**




### *High-Variability Accounts*

Figures 5-7 through 5-9 show the same plots for the weather-sensitive accounts with high variability. Note that a wider scale is used for the Theil's U plots than for the corresponding plots for the low-variability loads.

**Figure 5-7**
**Median Relative Hourly Error**
**Summer Uncurtailed Weather-Sensitive High-Variability Accounts**

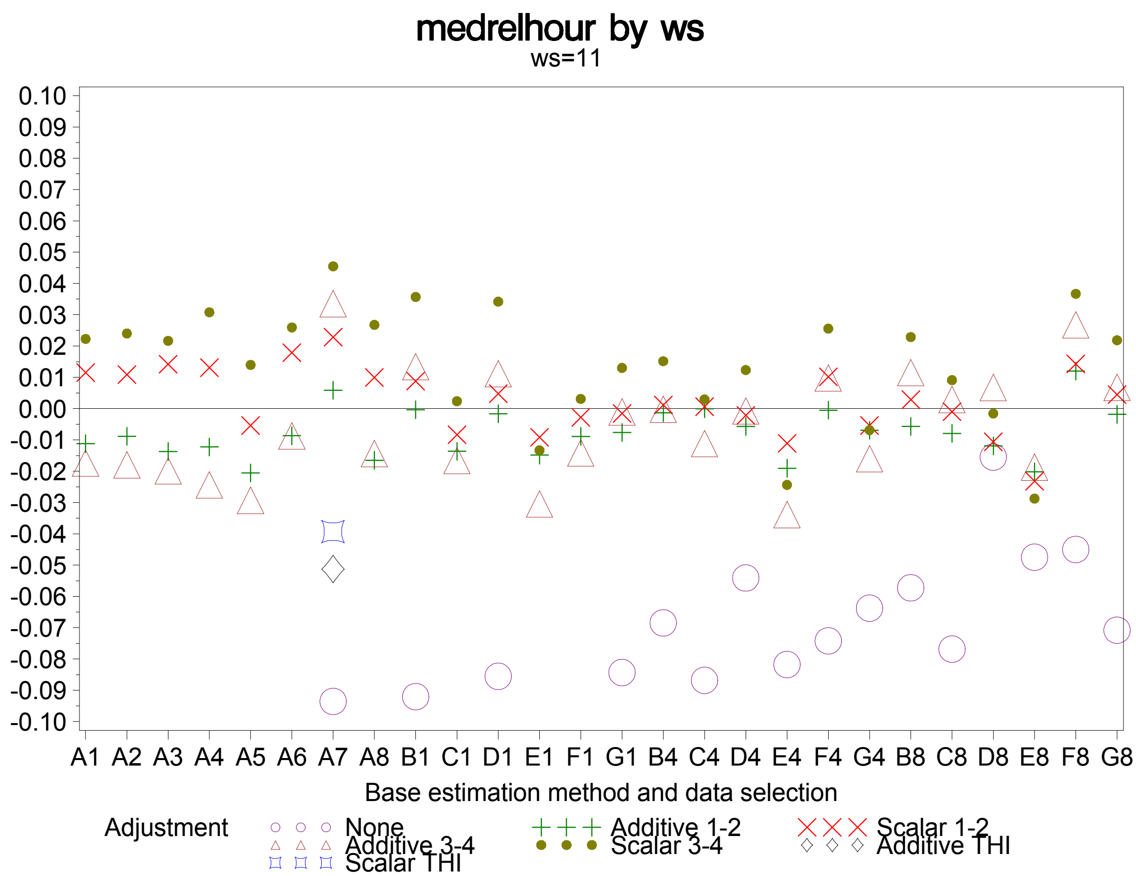


**Figure 5-8**
**Median Account Theil's U**
**Summer Uncurtailed Weather-Sensitive High-Variability Accounts**

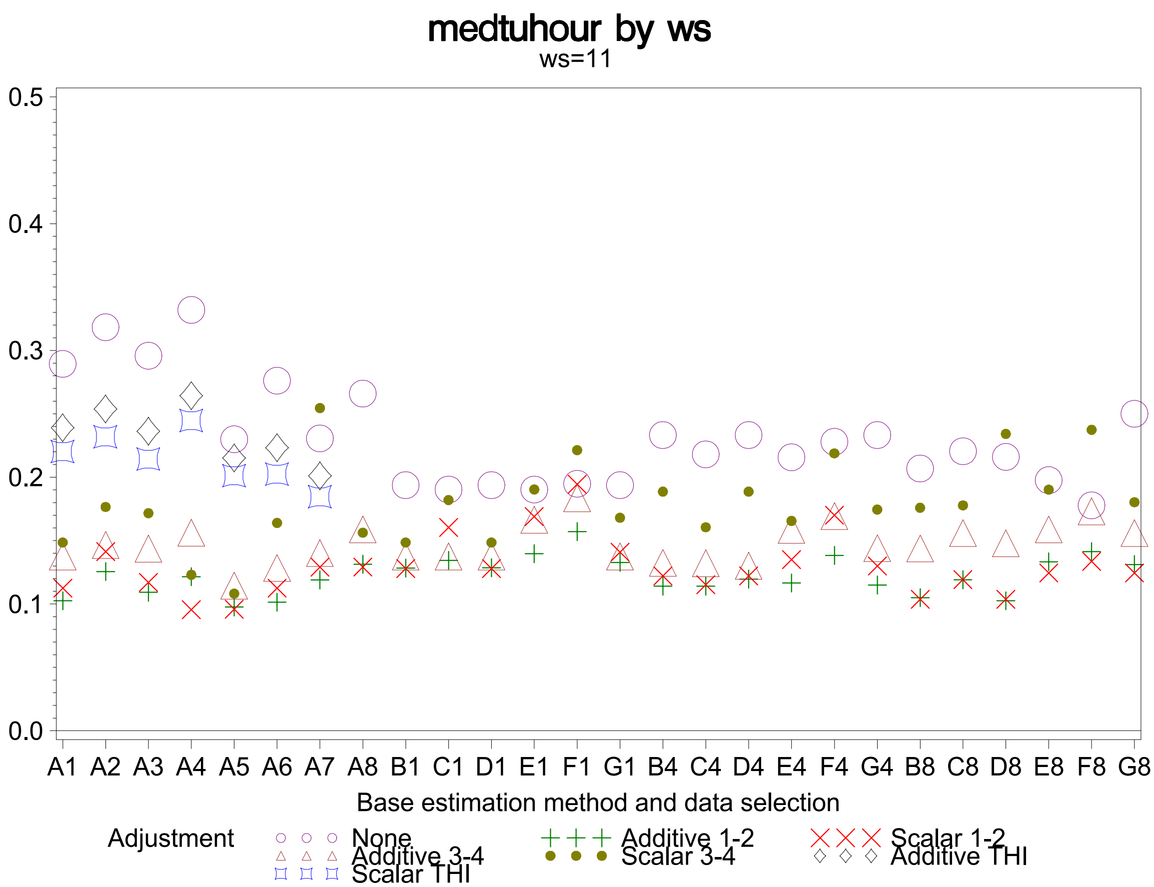
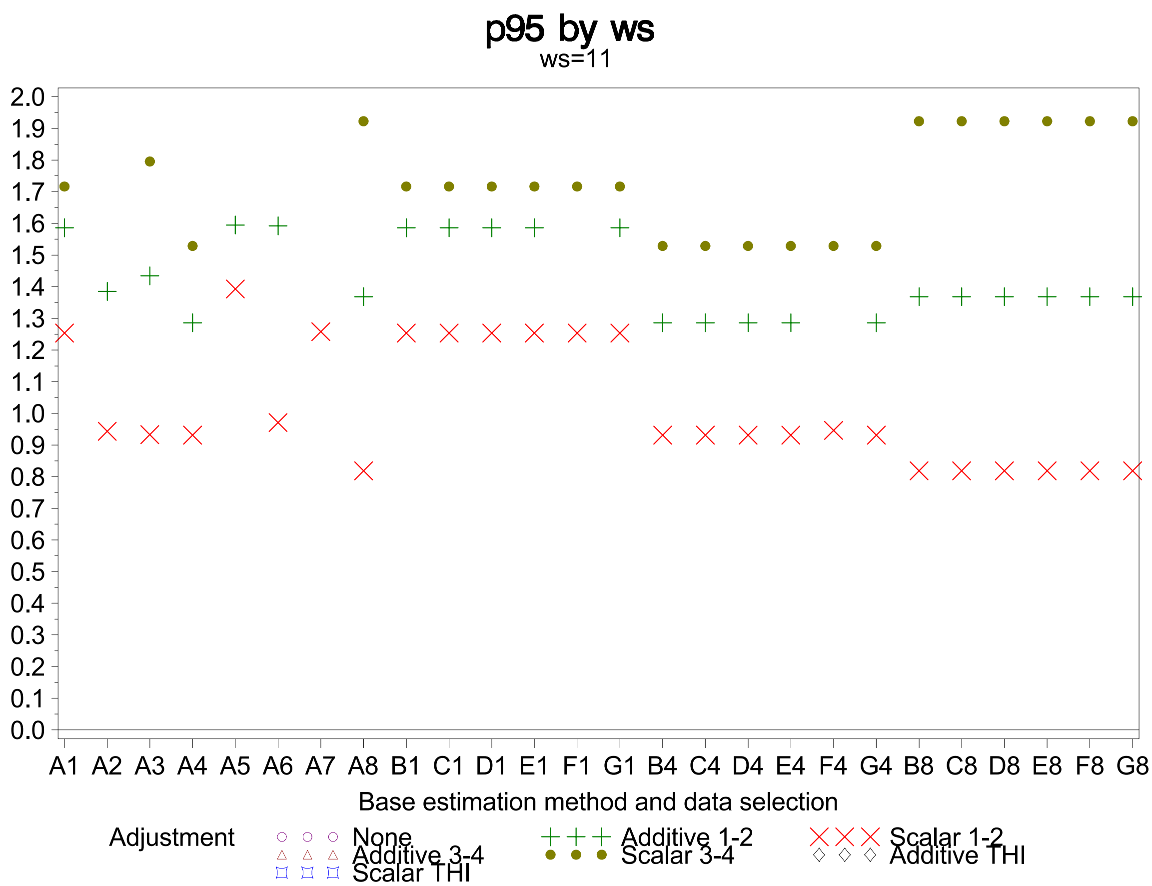


**Figure 5-9**
**95th Percentile Account Theil's U**
**Summer Uncurtailed Weather-Sensitive High-Variability Accounts**




Not surprisingly, these plots show that accounts with high variability are generally more difficult to predict accurately, even when they are weather-sensitive. Median relative errors are farther from zero for all methods, and are more than 10 percent off for the unadjusted averages. Median Theil's U is around 20 percent for the unadjusted weather models, and closer to 30 percent for the unadjusted averages. The additive adjustment to the last two hours still tends to minimize both bias and median Theil's U for each estimation-selection combination, though in some cases the scalar adjustment does somewhat better. In terms of the extreme Theil's U, the scalar adjustment always does best. For the averages, the THI adjustments are better than no adjustment, but worse than any of the adjustments to the pre-curtailment hours. Without adjustment, the weather models have smaller bias than the averages, but with adjustment to pre-curtailment hours the averages perform about as well as the weather models.

### 5.4.2 Non-Weather-Sensitive Loads

#### Low-Variability Accounts

Figures 5-10 through 5-12 show the performance for non-weather-sensitive low-variability accounts in the summer.

**Figure 5-10**
**Median Relative Hourly Error**
**Summer Uncurtailed Non-Weather-Sensitive Low-Variability Accounts**

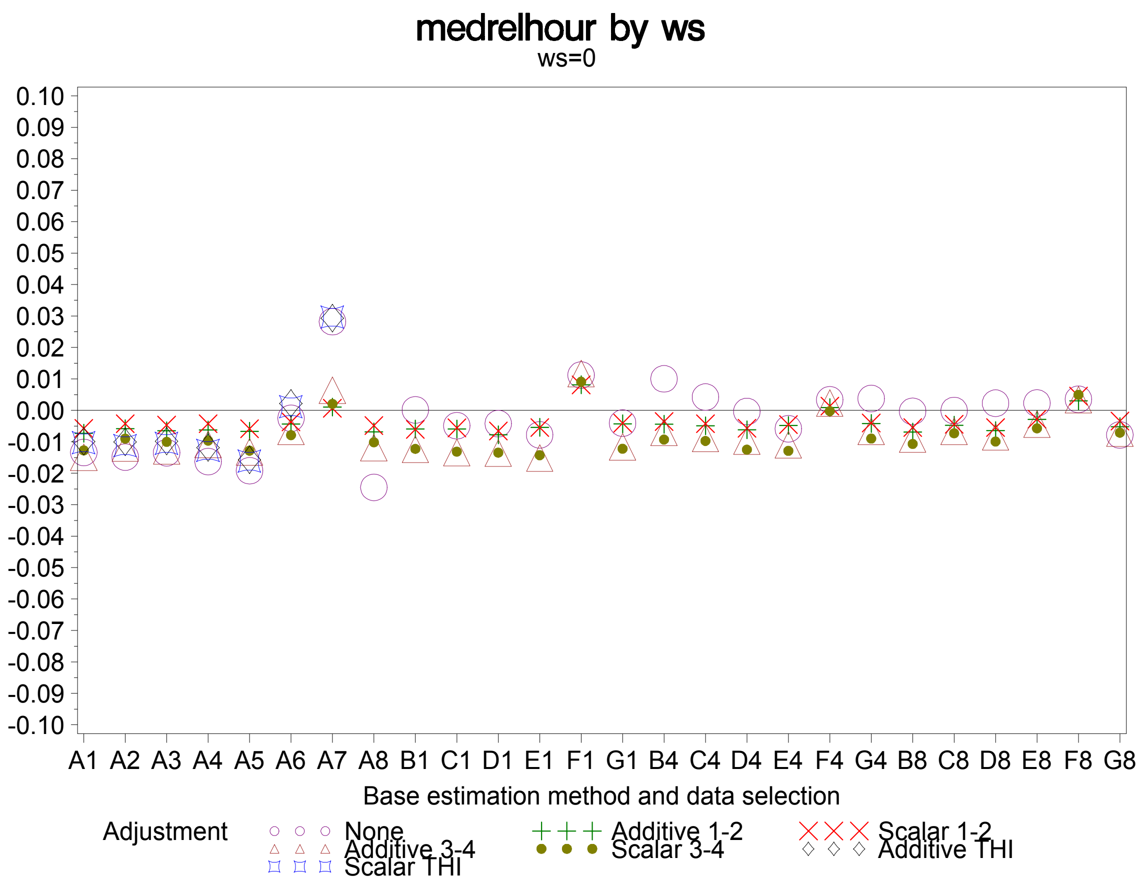


**Figure 5-11**
**Median Account Theil's U**
**Summer Uncurtailed Non-Weather-Sensitive Low-Variability Accounts**

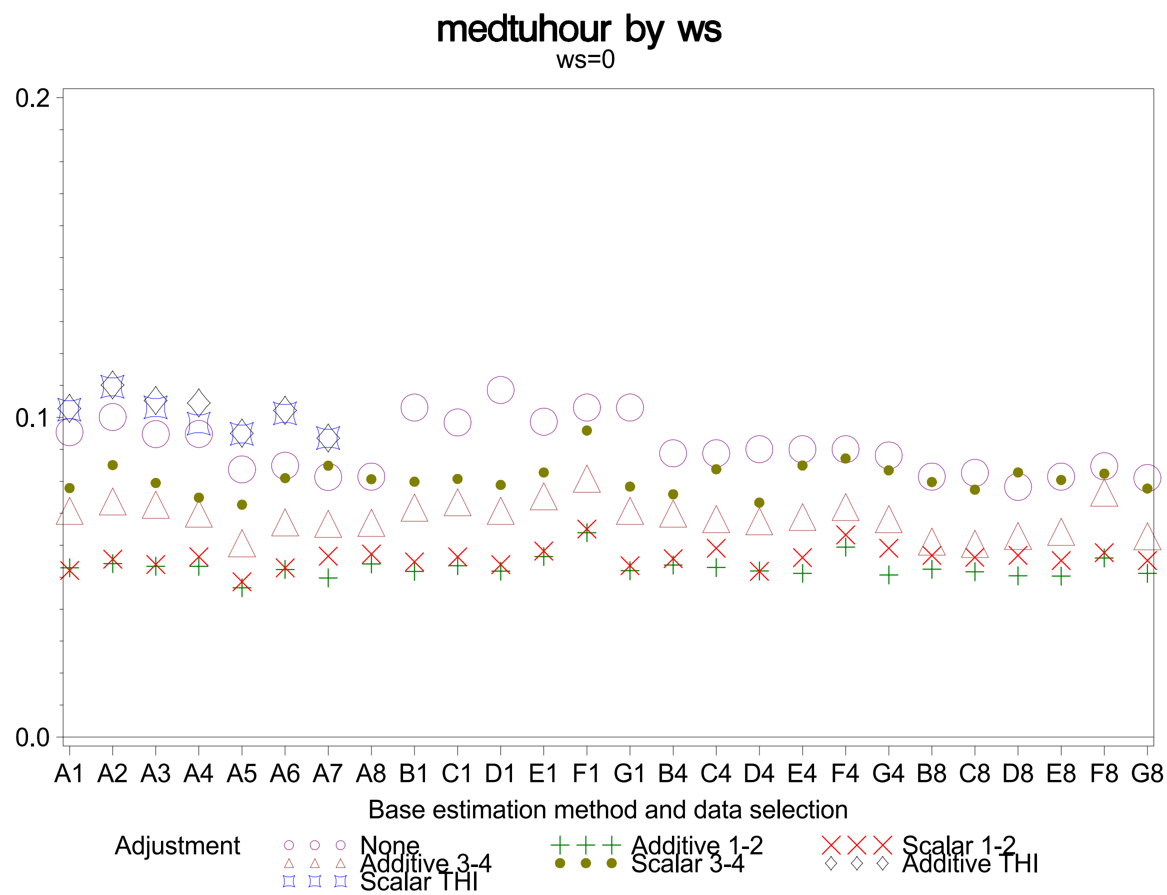
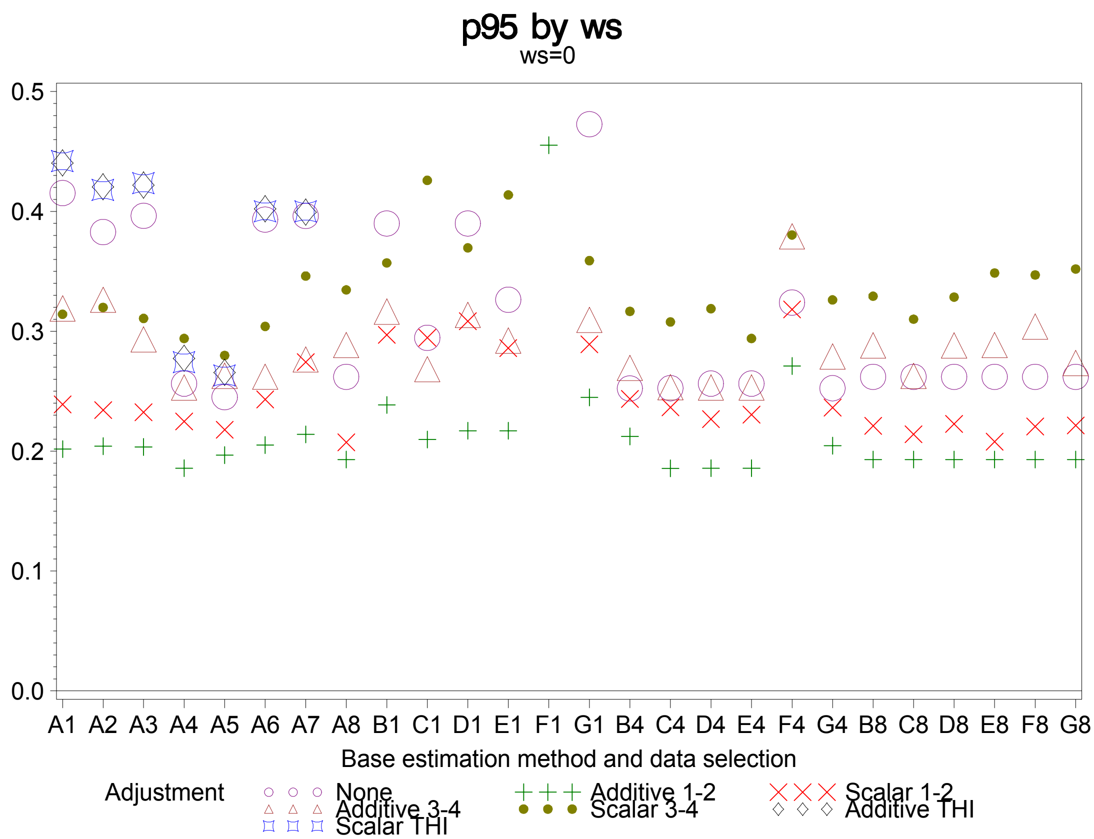


**Figure 5-12**
**95th Percentile Account Theil's U**
**Summer Uncurtailed Non-Weather-Sensitive Low-Variability Accounts**




Observations from these plots include the following.

1. Most of the unadjusted averages (A1 through A5) show some negative bias (around –1 percent) even for these non-weather-sensitive accounts. The reason may be that these accounts do have some weather-sensitive load components, but the diagnostics classify the account as non-weather-sensitive because other variations generally dominate the weather-sensitive load components.
2. The unadjusted High 5 average (A7) tends to over-adjust, with a median +3 percent relative error.
3. With adjustments, all the methods come within $\pm$ 1 percent median relative error.
4. For all estimation-selection methods, the adjustment to the last two hours again has the lowest median Theil's U of all the adjustment methods, and in most cases has the smallest or nearly smallest median relative hourly error.
5. The THI adjustments have the worst bias and worst median Theil's U for all the averages.
6. The conditional weather models perform about the same as the averages. However, the lag models have somewhat worse variability indicated by higher median Theil's U. For

weather models based on less than a full season, the extremes (indicated by 95th percentile Theil's U) are much worse with the lag model.

### *High-Variability Accounts*

The final set of figures for tests of uncurtailed accounts are for the non-weather-sensitive high-variability accounts. As for the weather-sensitive accounts, the high-variability accounts show median relative errors farther from zero and larger median and 95th percentiles of account Theil's U. That is, both systematic errors and overall error magnitudes are larger than for the low-variability accounts.

For each adjustment method, the full-season weather models have slightly less bias but much worse variability than the averages, indicated by higher median and 95th percentile Theil's U. Thus, for high-variability accounts without weather drivers, using a closer set of days is more helpful than having a longer data series. The 10- and 20-day weather models perform comparably to the averages. The additive adjustment to the last two hours is still the best performer in terms of Theil's U and median relative error. However, the other adjustments, other than the THI, are nearly as good.

**Figure 5-13**
**Median Relative Hourly Error**
**Summer Uncurtailed Non-Weather-Sensitive High-Variability Accounts**

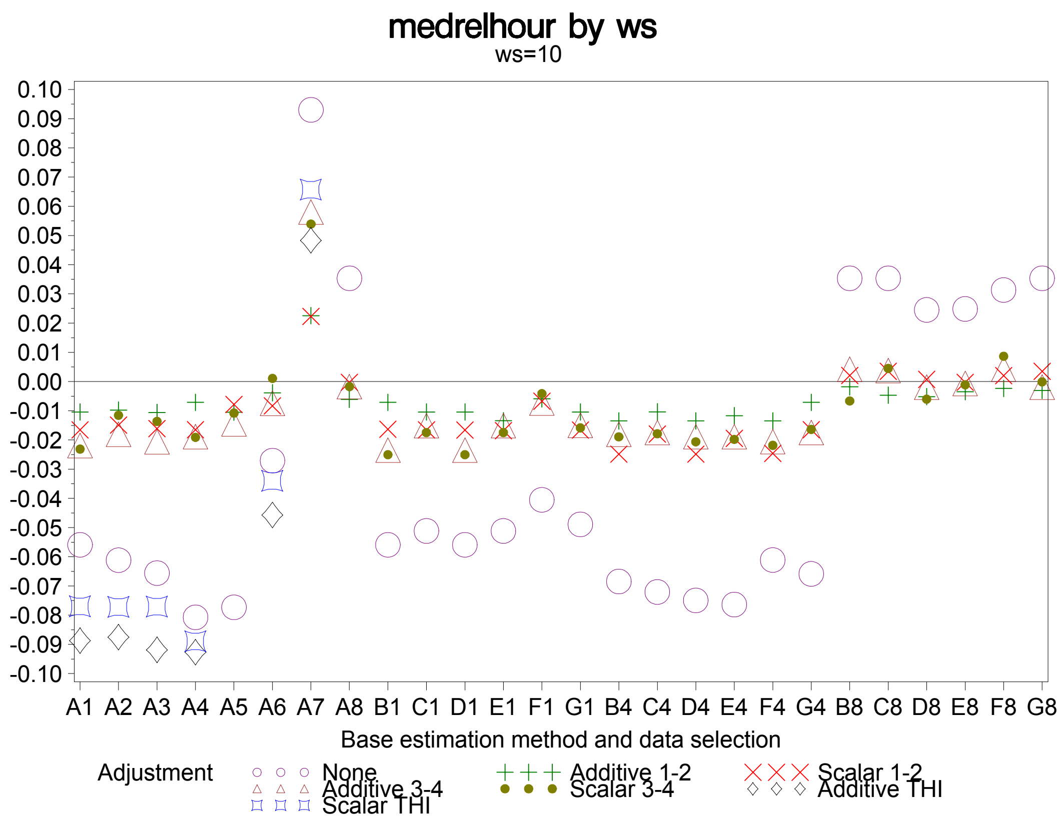


**Figure 5-14**
**Median Account Theil's U**
**Summer Uncurtailed Non-Weather-Sensitive High-Variability Accounts**

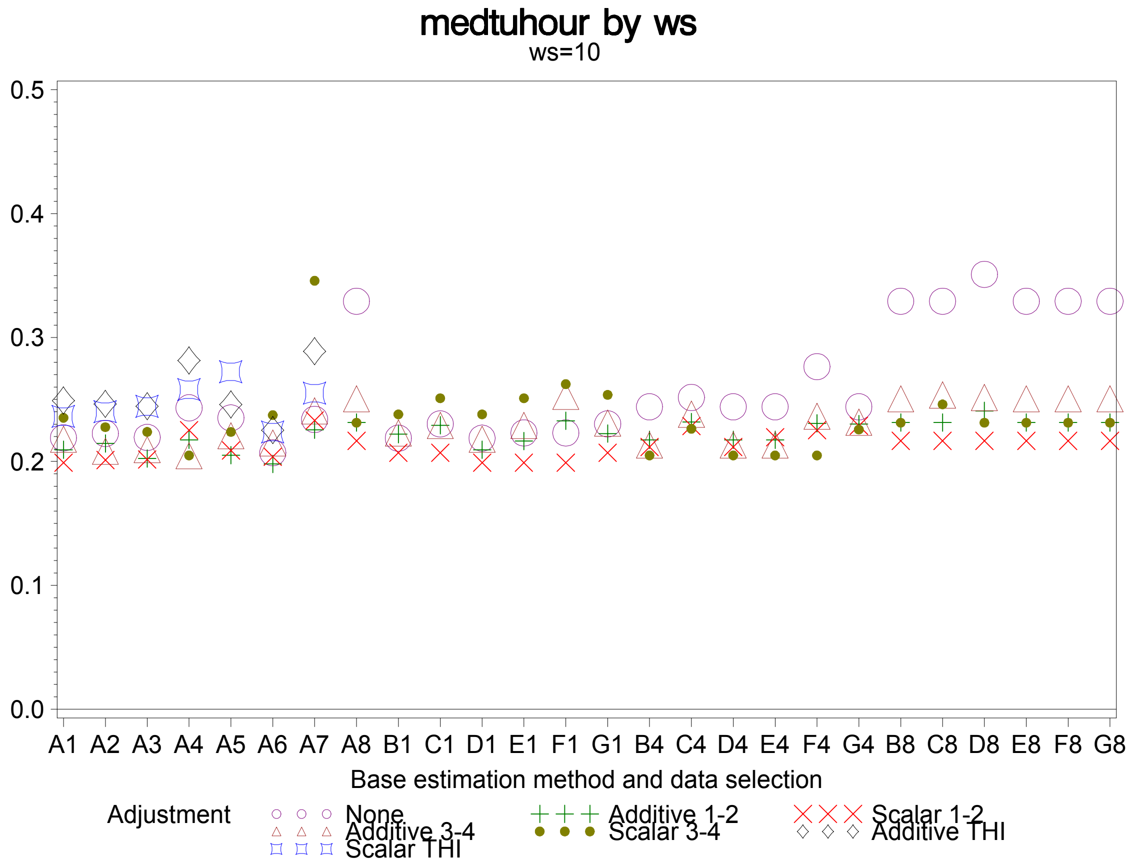
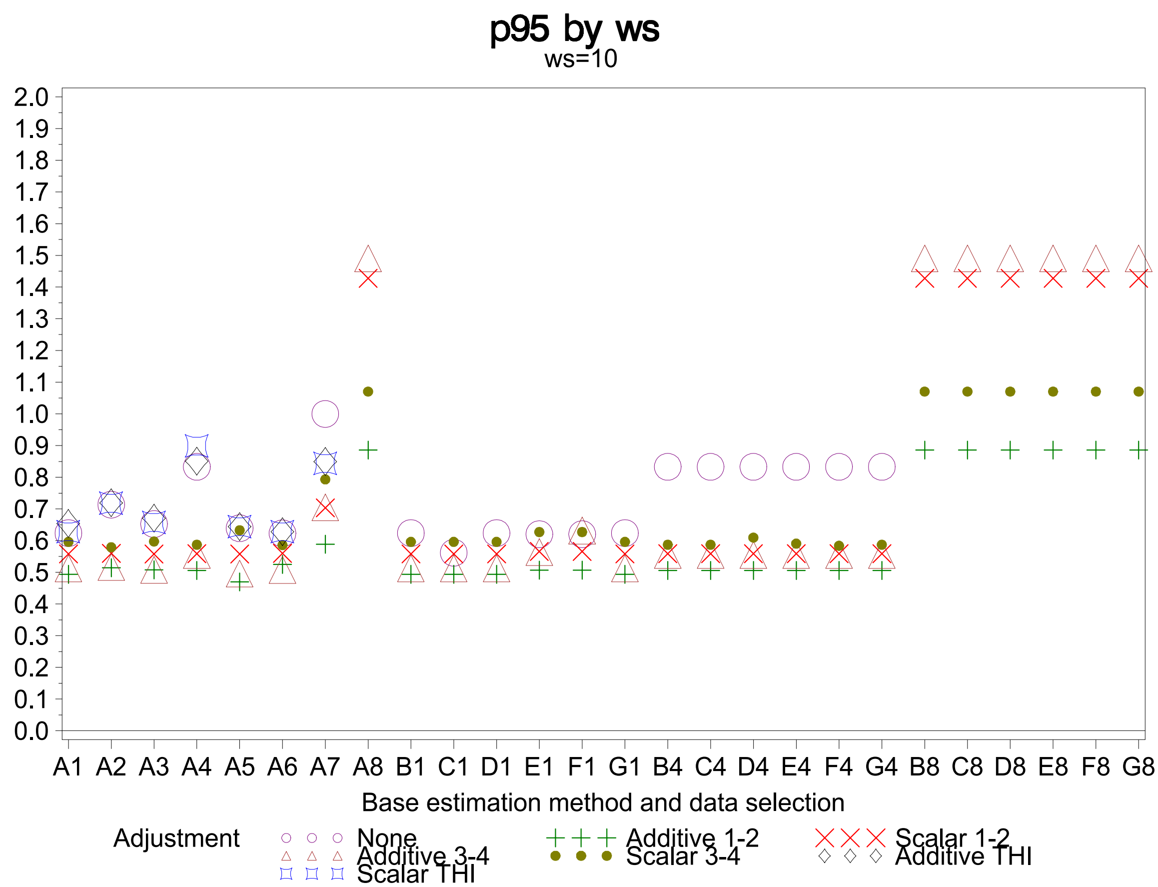


**Figure 5-15**
**95th Percentile Account Theil's U**
**Summer Uncurtailed Non-Weather-Sensitive High-Variability Accounts**




### 5.4.3 *Summary*

Based on the results for uncurtailed accounts above, it appears that the following can be recommended for summer baselines:

1. Additive adjustment to the two hours before curtailment appears to be the best adjustment, in terms of both bias and overall error magnitude, for all methods, for weather-sensitive and non-weather-sensitive, high- and low-variability accounts.

2. For weather-sensitive accounts, weather models perform somewhat better than averages. However, adjusted averages can perform nearly as well. Longer input data series improves weather model performance, but only slightly.

3. For non-weather-sensitive accounts, conditional weather models (i.e., models that drop inappropriate weather variables based on diagnostic screening) do not increase variability compared to simpler averages. However, for high-variability non-weather-sensitive loads

full-season weather models can be higher variability than either simple averages or weather models based on 10 or 20 days.

4. The High 5 of 10 selection rule (highest 5 of last 10) gives the best unadjusted average for weather-sensitive loads. However, with additive adjustment to the last two hours the method is no better than the other averages. For non-weather-sensitive accounts the High 5 of 10 average has positive bias unless an adjustment is included.
5. The THI adjustments have higher variability and worse bias than the other adjustments to averages.

## 5.5 Uncurtailed Accounts, Nonsummer

Results for weather-sensitive accounts in the nonsummer period corresponding to those shown above are shown in the next series of figures. Observations for all nonsummer results are provided at the end.

### *5.5.1 Weather-Sensitive Accounts*

#### *Low-Variability Accounts*

**Figure 5-16**
**Median Relative Hourly Error**
**Nonsummer Uncurtailed Weather-Sensitive Low-Variability Accounts**

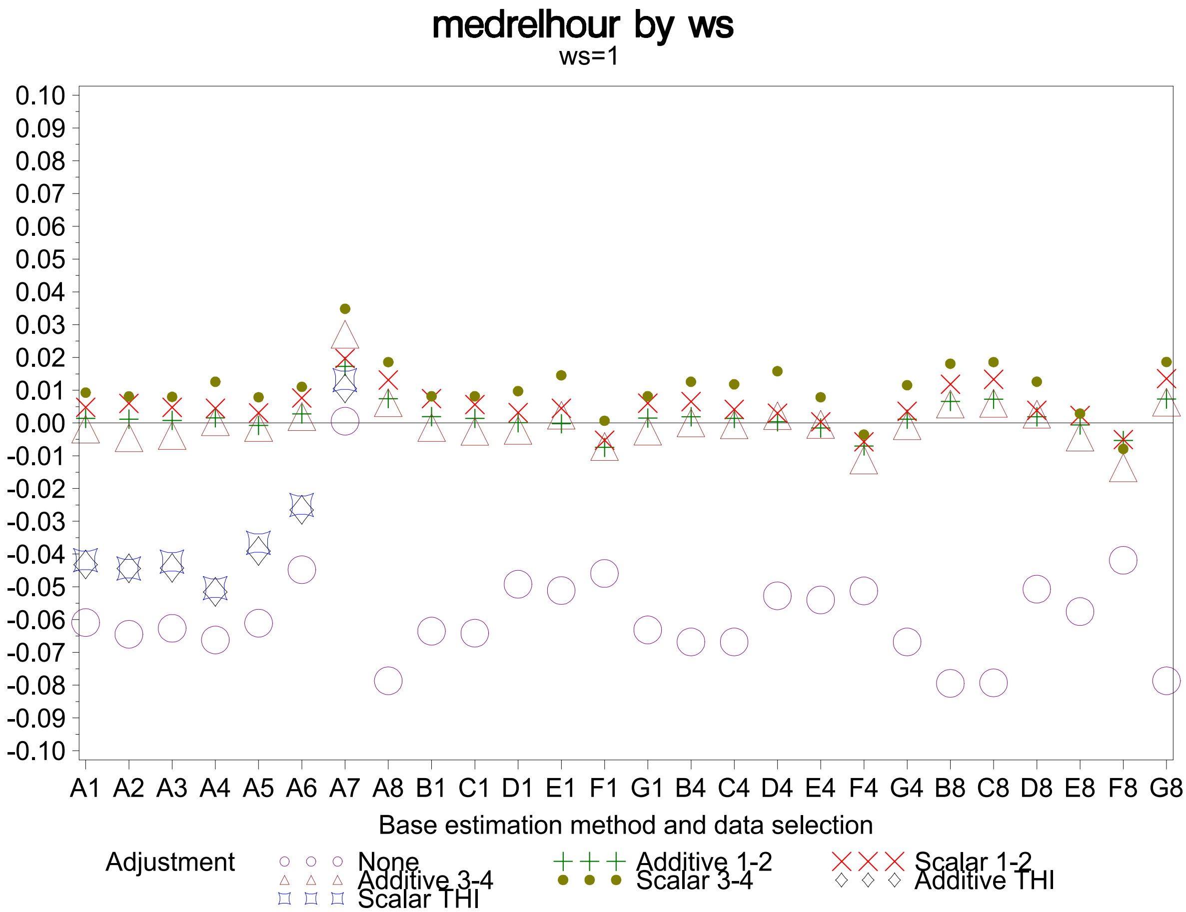


**Figure 5-17**
**Median Account Theil's U**
**Nonsummer Uncurtailed Weather-Sensitive Low-Variability Accounts**

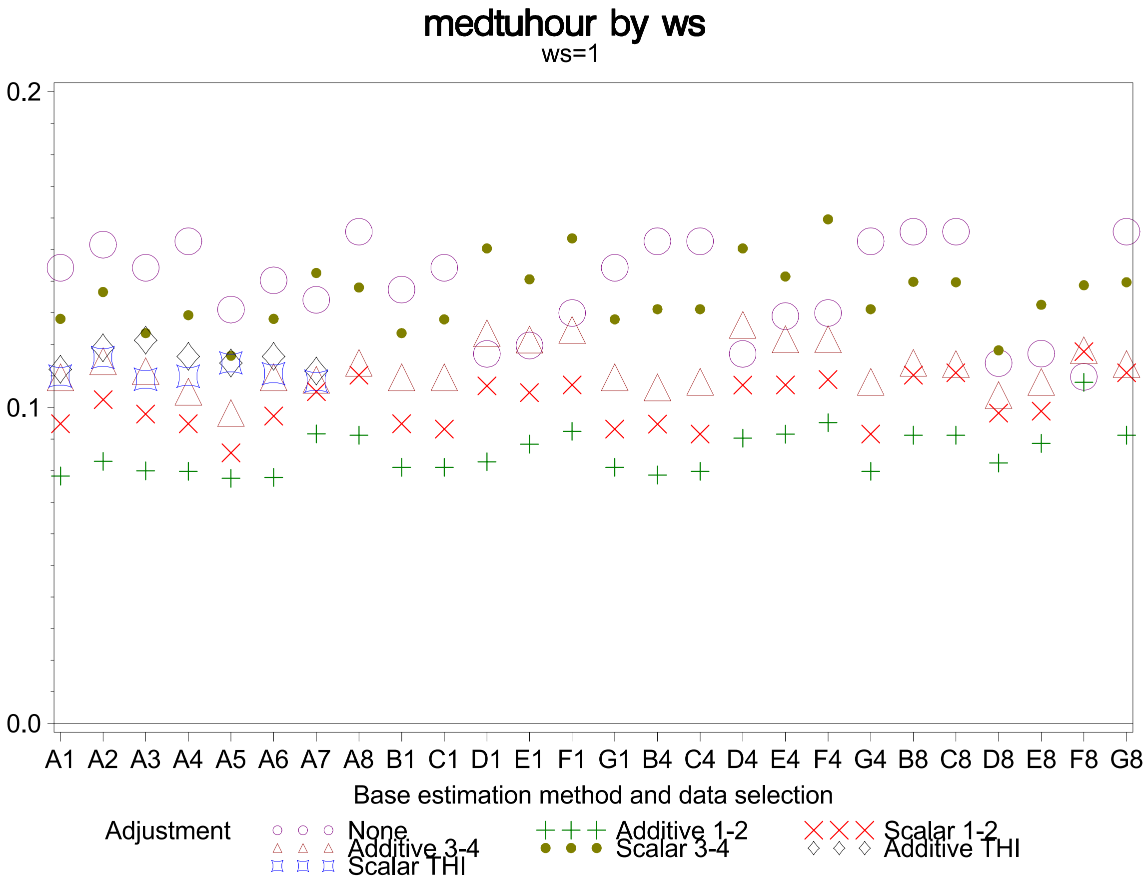
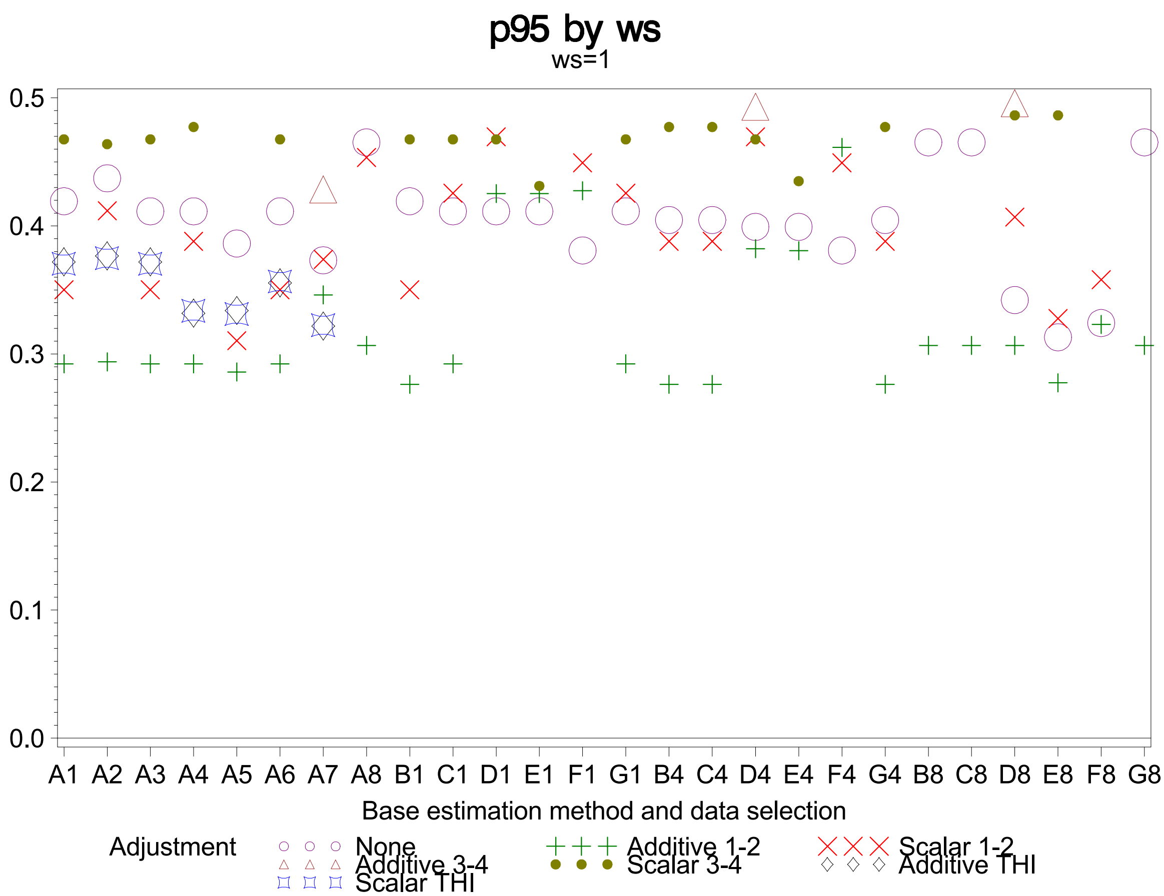


**Figure 5-18**
**95th Percentile Account Theil's U**
**Nonsummer Uncurtailed Weather-Sensitive Low-Variability Accounts**



### *High-Variability Accounts*

**Figure 5-19**
**Median Relative Hourly Error**
**Nonsummer Uncurtailed Weather-Sensitive High-Variability Accounts**

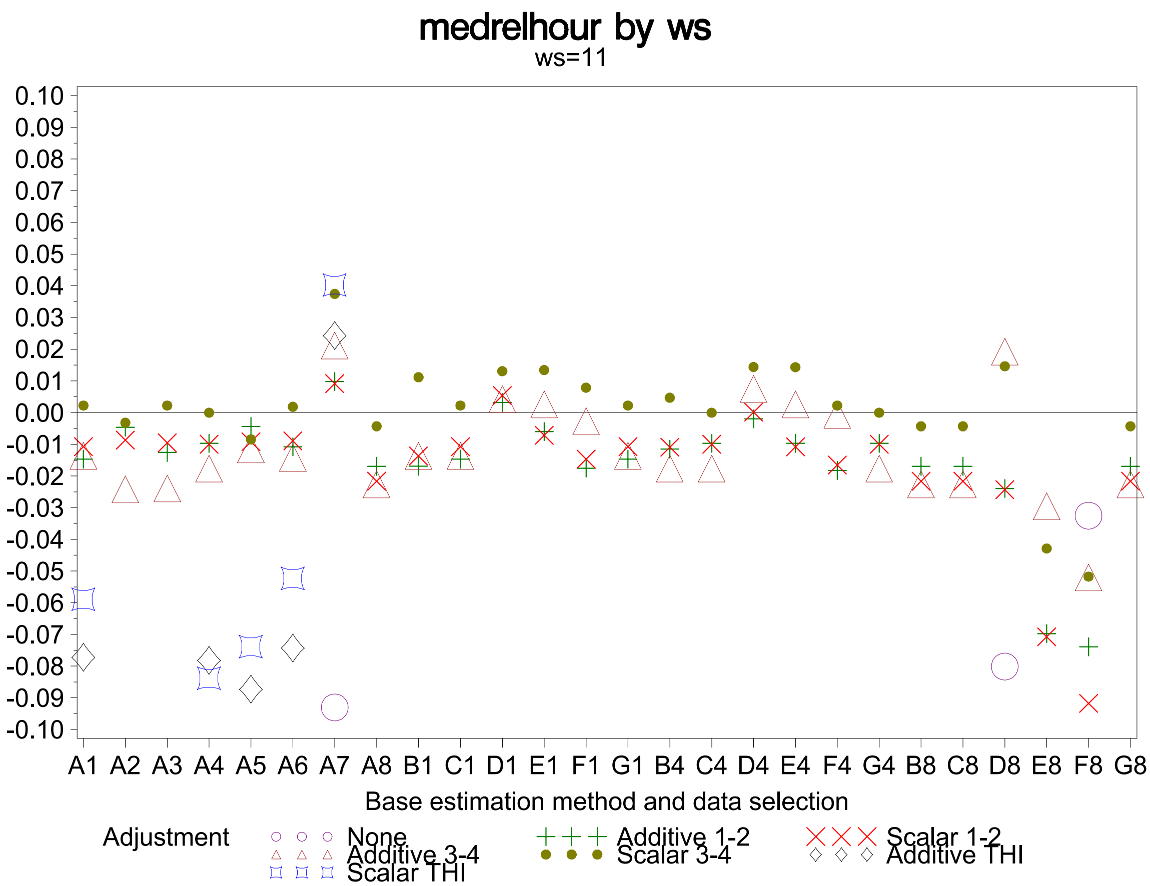


**Figure 5-20**
**Median Account Theil's U**
**Nonsummer Uncurtailed Weather-Sensitive High-Variability Accounts**

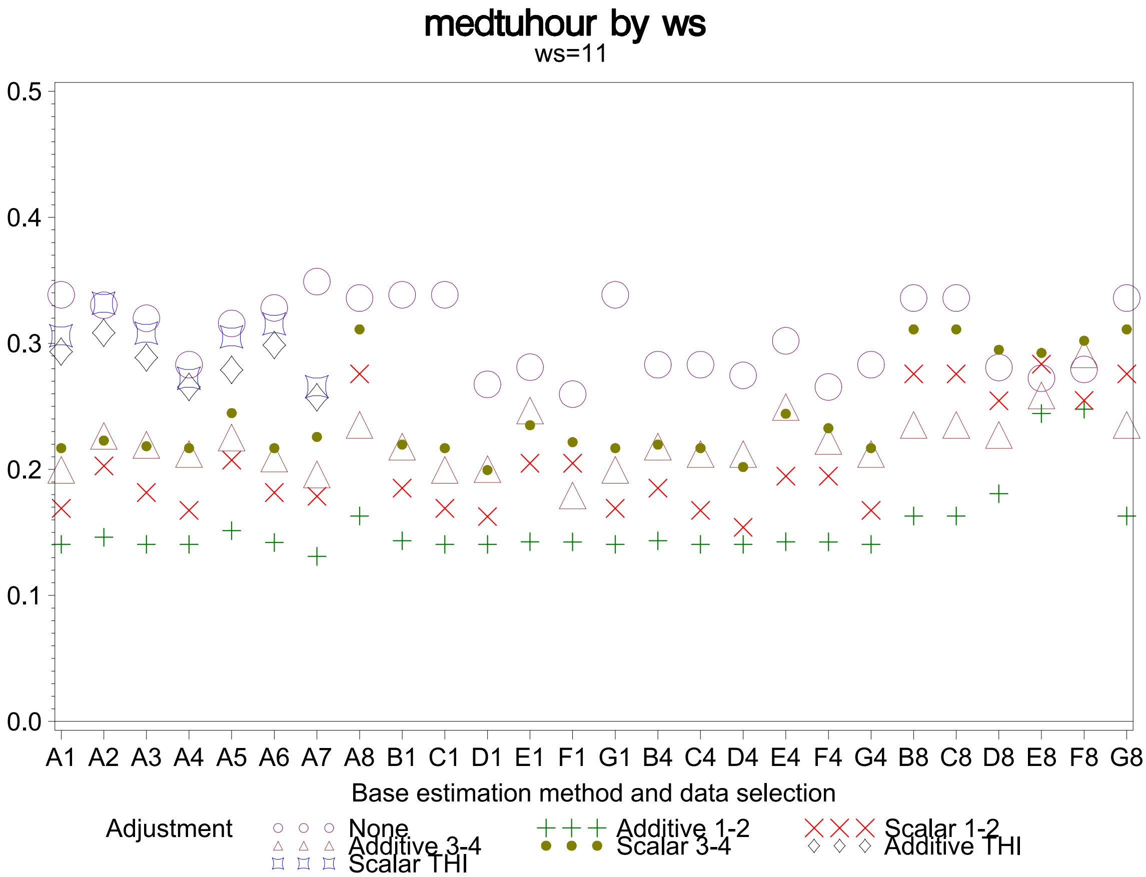


**Figure 5-21**
**95th Percentile Account Theil's U**
**Nonsummer Uncurtailed Weather-Sensitive High-Variability Accounts**

### 5.5.2 Non-Weather-Sensitive Accounts

#### Low-Variability Accounts

**Figure 5-22**
**Median Relative Hourly Error**
**Nonsummer Uncurtailed Non-Weather-Sensitive Low-Variability Accounts**

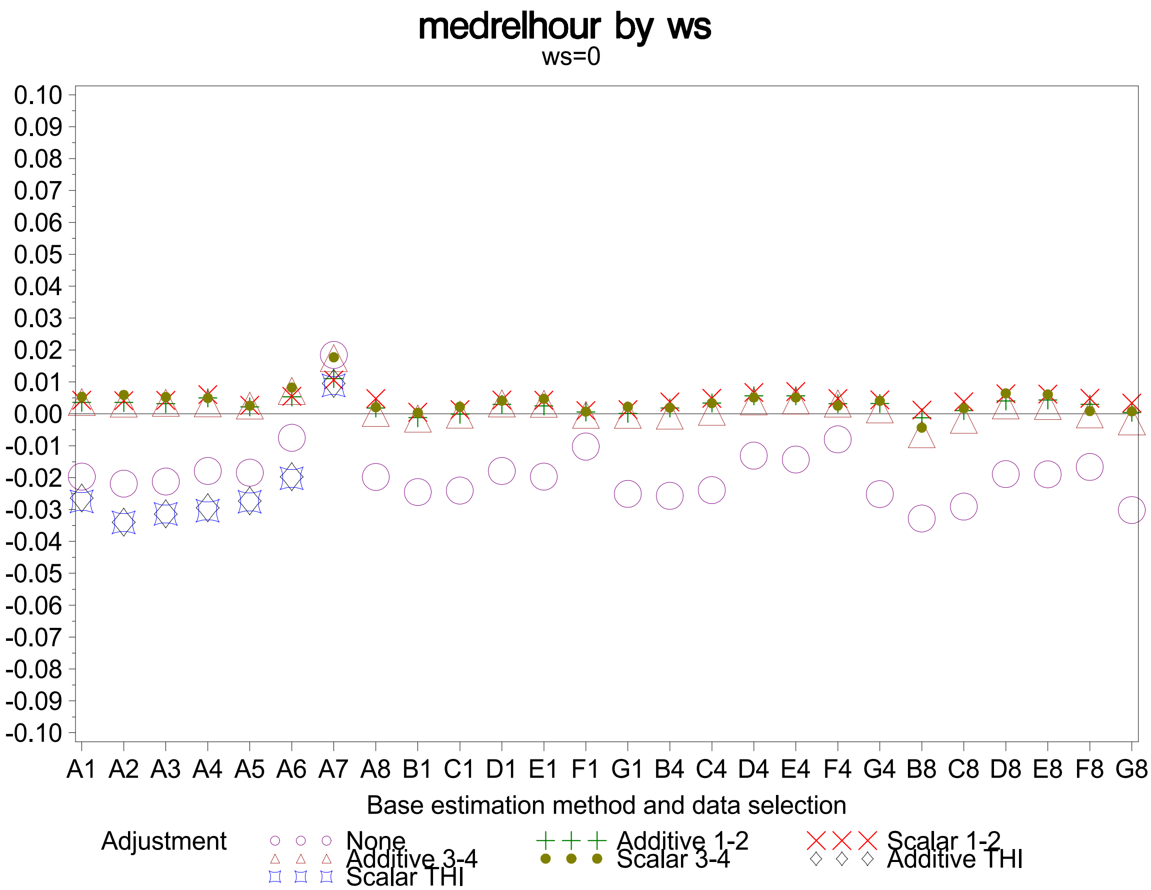


**Figure 5-23**
**Median Account Theil's U**
**Nonsummer Uncurtailed Non-Weather-Sensitive Low-Variability Accounts**

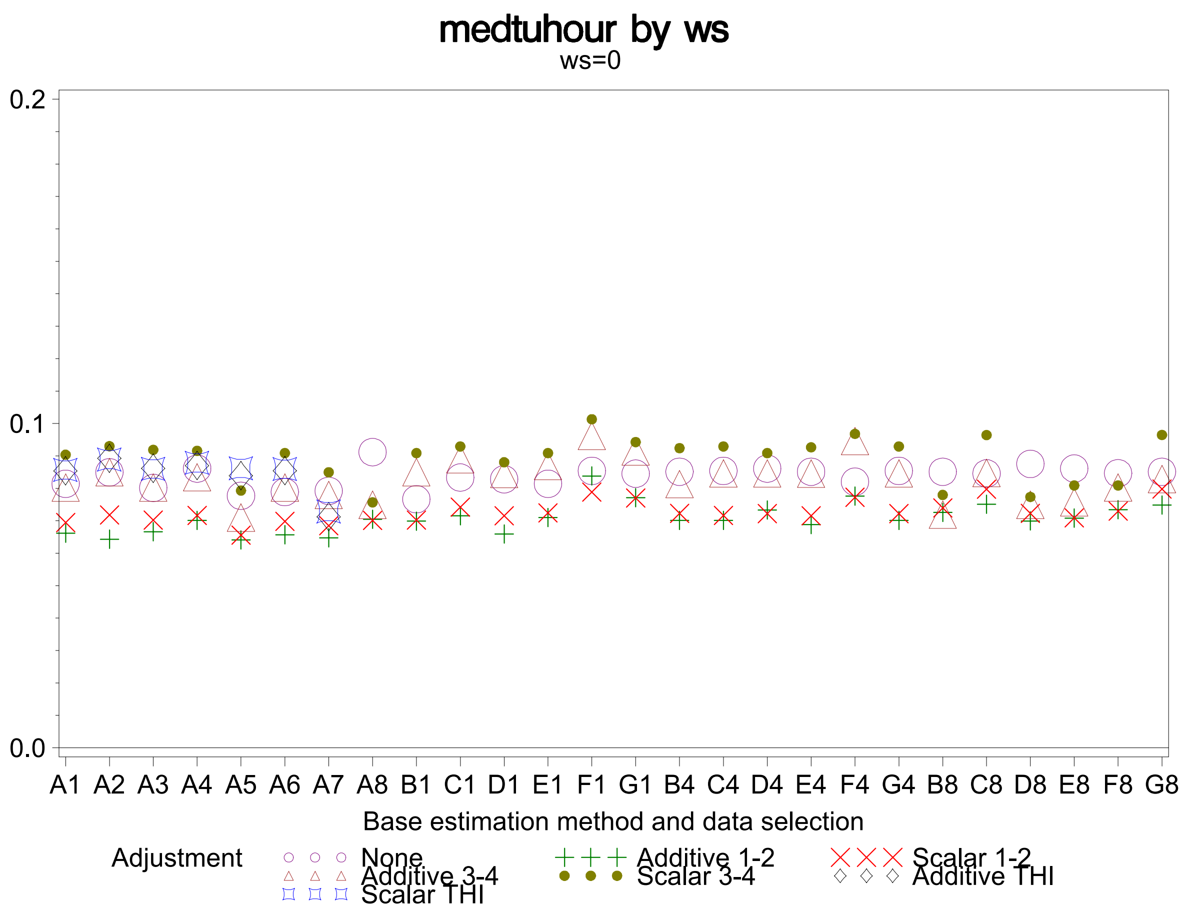
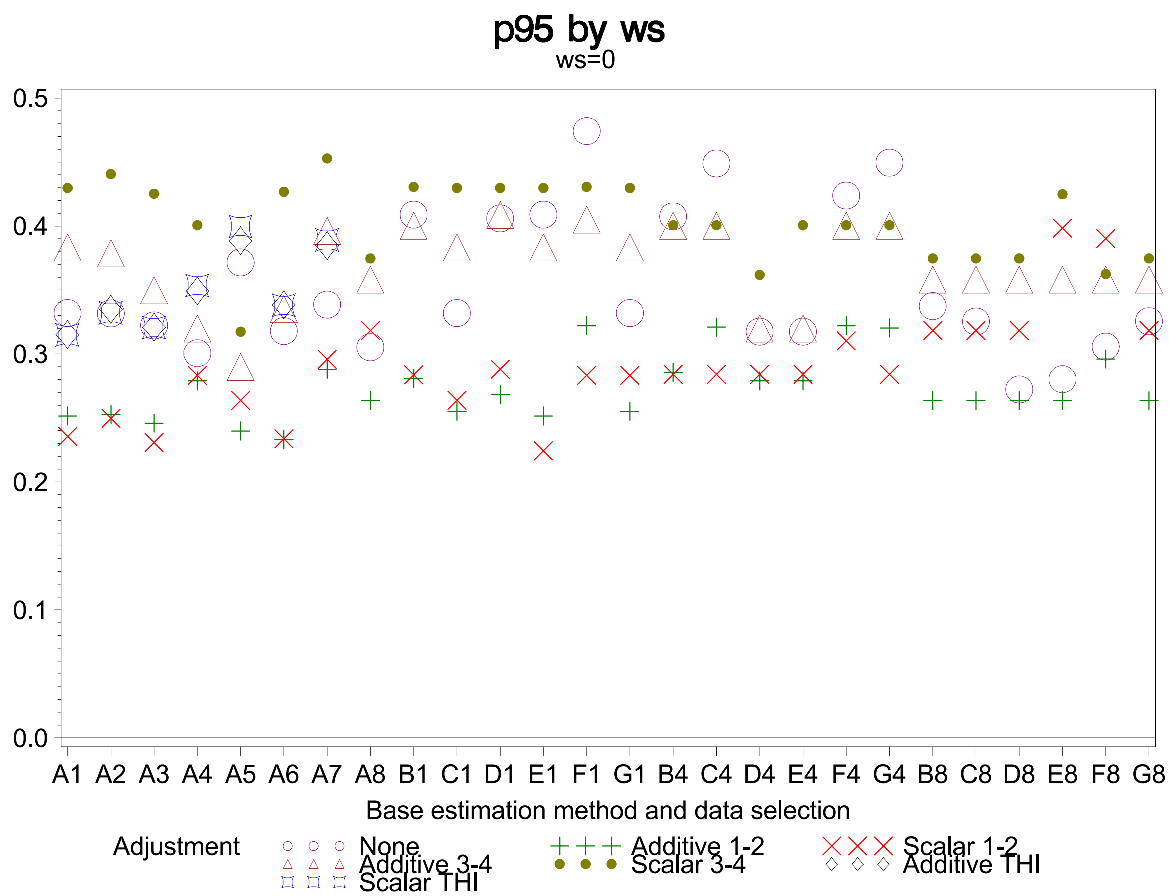


**Figure 5-24**
**95th Percentile Account Theil's U**
**Nonsummer Uncurtailed Non-Weather-Sensitive Low-Variability Accounts**



### *High-Variability Accounts*

**Figure 5-25**
**Median Relative Hourly Error**
**Nonsummer Uncurtailed Non-Weather-Sensitive High-Variability Accounts**

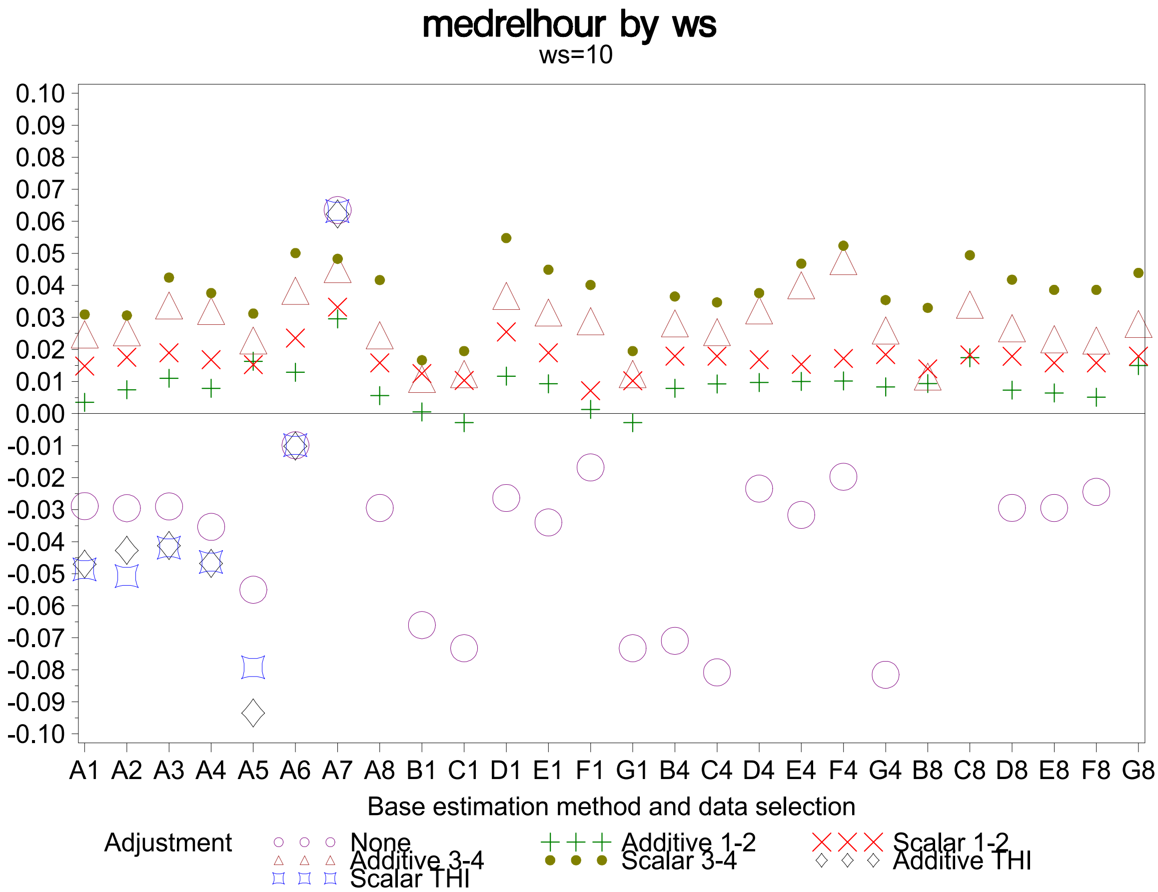


**Figure 5-26**
**Median Account Theil's U**
**Nonsummer Uncurtailed Non-Weather-Sensitive High-Variability Accounts**

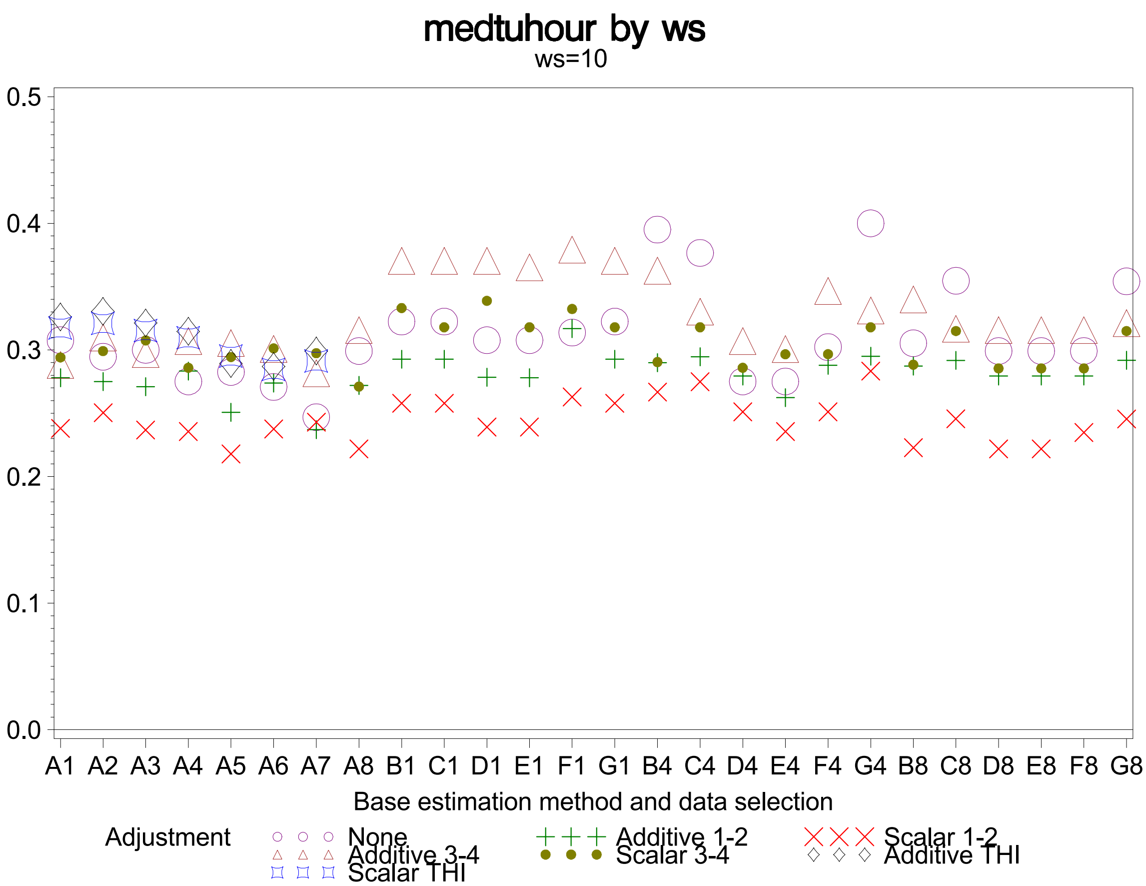
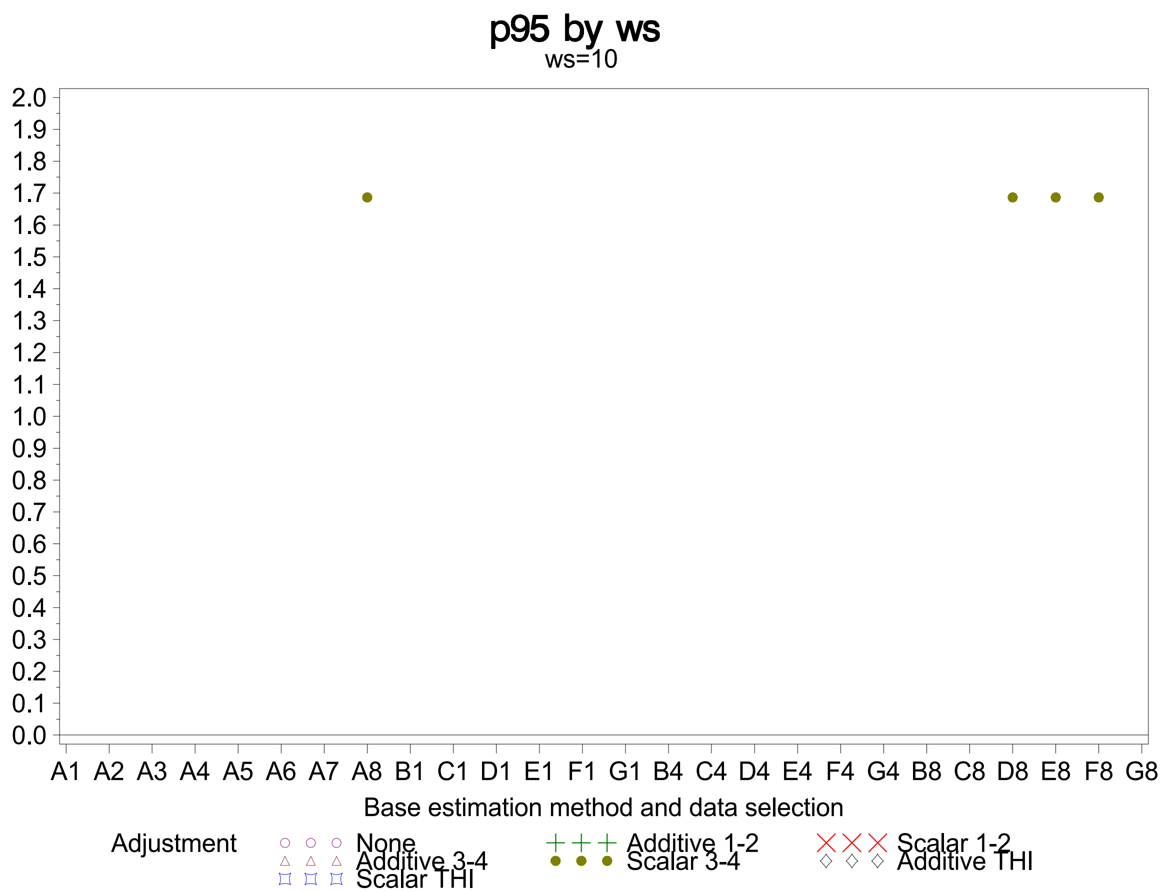


**Figure 5-27**
**95th Percentile Account Theil's U**
**Nonsummer Uncurtailed Non-Weather-Sensitive High-Variability Accounts**




### 5.5.3 Observations for Nonsummer Accounts

Observations for the uncurtailed nonsummer accounts are as follows:

- Accurate weather modeling is more difficult than for summer loads. The performance of weather models is not clearly better than that of the averages even for low-variability weather-sensitive accounts. For high-variability weather-sensitive accounts the adjusted averages do better in terms of both bias and overall error magnitude.
- For low-variability weather-sensitive accounts, the unadjusted High 5 of 10 average performs best in terms of small bias and low variability. For high-variability and/or non-weather-sensitive accounts, the unadjusted High 5 method does not perform as well as adjusted methods.
- For all estimation-selection rules and three of the four load types, the additive adjustment to the last two hours has smallest median Theil's U, and smallest or close to smallest bias.

However, for high-variability non-weather-sensitive accounts, the scalar adjustment has slightly worse bias but somewhat lower Theil's U.

- For non-weather-sensitive loads there is little difference across estimation methods and selection rules in either bias or overall error magnitude.
- Unadjusted averages and weather models tend to understate baselines.

## 5.6 CURTAILED SUMMER ACCOUNTS

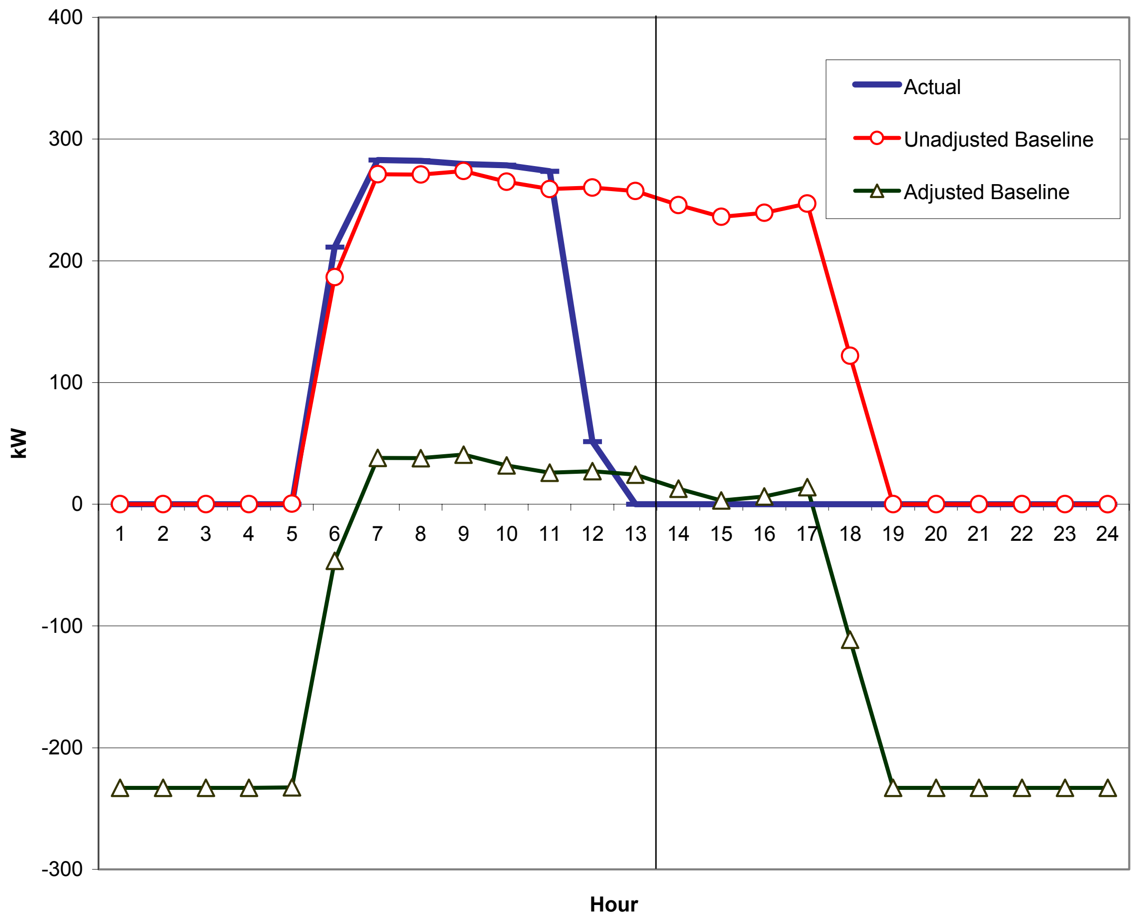
### 5.6.1 Reference Case for Comparisons

To provide similar method comparisons to those presented above for actual curtailment events, it was necessary to select a method to represent the best estimate of the true load that would have occurred in the absence of curtailment. Based on the results for the uncurtailed load, we initially selected the conditional full-season degree-day model with additive adjustment to the last two hours.

However, the comparative results immediately illustrated a key shortcoming of this method in actual curtailment situations. Many operations will send a shift home early or cancel a shift entirely when faced with a curtailment notice or the likely prospect of one. When this happens, the load in the two hours before curtailment is near zero, and the adjustment of the average or modeled load to these two hours produces a negative or zero baseline for the duration of the curtailment period.

An example of such an account is shown in the next figure. Although the actual load was reduced to zero in hour 13, the curtailment did not begin until hour 14. The adjusted baseline was near zero during usual operating hours, and became substantially negative after the hours when load customarily dropped to zero.

**Figure 5-28**
**Adjustment to Last Two Hours for Account That Curtailed before Required Time**




To avoid this type of problem, we defined the "true" load for curtailed cases as the unadjusted full-season daily degree-day model.

### *5.6.2 Findings for Curtailed Accounts*

#### *Errors in Baseline Estimation*

The displays for the curtailed accounts are shown in Appendix B. With the true load in the absence of curtailment defined as indicated above, the results for the curtailed cases generally followed similar patterns to those seen for the uncurtailed cases, with the following differences.

1. All the full-season weather models show very little bias and low variability. This is because there is little difference among the models, as indicated by the uncurtailed results. Thus, the relative errors, defined as the relative difference between a method and

the reference method, which is one of the full season models, are small. For the model chosen as the reference (unadjusted D8), all the errors are necessarily 0.

2. The unadjusted methods tend to show smaller biases and smaller Theil's U than for the same load types with uncurtailed accounts. This result makes sense given the similarity between the weather models and the averages, with a weather model defined as "truth." That is, the errors are reduced because part of the natural variability in actual loads is eliminated by defining truth as the result of the weather model.
3. The adjustment to the last two hours, rather than looking like the best method in most cases, shows substantial negative bias, and large overall error magnitude.
4. The adjustment to the third and fourth hours prior to curtailment generally has greater negative bias and worse overall error magnitude than for the same load type with uncurtailed accounts. This result indicates that some of the "pre-event" curtailment effects are seen in these hours as well as in the immediate last two hours.

Thus, the methods that work best in the abstract case of loads that were not actually subject to a curtailment process can be very problematic in practice. The recommendations in Section 6 address this problem.

### *Errors in Estimating Demand Reduction*

In addition to looking at errors relative to estimated actual loads, we also looked at errors relative to the estimated load reduction, as noted in Section 4. These results are also included in Appendix B.

In terms of program operations, the important question is how far off the estimated reduction is from the customer's actual load shed. While the magnitudes are greater, the qualitative patterns of errors relative to reduction amounts are similar to those for errors relative to load levels. Very roughly, the medians and Theil's U statistics for errors relative to estimated reductions are about three times as large as the errors relative to estimated load. This relationship corresponds to reductions being on the order of one-third of total load. Of course, for an individual account and hour the reduction may be anywhere from a small fraction to 100 percent of the load.

# 6 FINDINGS AND RECOMMENDATIONS

## 6.1 GENERAL FINDINGS

Important considerations in defining a baseline calculation method include:

- simplicity of calculation
- minimizing burden on participants and operators, in terms of costs, ease of understanding, and ease of operation
- limiting the potential for gaming
- limiting the potential for distortion if load is curtailed early
- ability to know the baseline immediately after a curtailment
- ability to know the baseline before making a curtailment decision
- minimizing method bias: systematic tendency to over- or under-state
- minimizing method variability: tendency to wide swings in estimates.

Specific concerns that have been addressed by developers of existing baseline methods for demand response programs include:

- providing accurate baselines for weather-sensitive accounts
- avoiding windfall credits for cool weather
- avoiding windfall credits for planned shut-downs.

Advantages and disadvantages of key method features in terms of these concerns are summarized in the table below. This table expands the qualitative summary provided in Section 3 with findings from the test results presented in Section 5.

**Table 6-1**
**Advantages and Disadvantages of Key Baseline Method Features**
**Based on Qualitative Considerations and Test Results**

| Baseline Method | Variant | Pros | Cons |
|---|---|---|---|
| Average | Any | Simple, easy to use and understand, low cost | Tends to understate baseline for weather-sensitive loads, especially if unadjusted |
| | High 5 of last 10 days | Partial adjustment for weather-sensitive loads | Still tends to understate baseline for weather-sensitive loads |
| | | | Can allow windfall load reduction credit on cool days |
| Regression | Any | Provides baseline corresponding to particular weather conditions of curtailment day | More complex, harder to understand, higher cost |
| | | | If observations don't include conditions as extreme as the curtailment day, model estimate may be inaccurate |
| | | | If account isn't weather-sensitive, may be less accurate than simpler methods |
| | Full Season | Adequate data and range of variation to yield accurate coefficients | Operating conditions from the period data are taken from may be different from curtailment day |
| | Recent 10 days | Operating conditions more likely to be similar to curtailment day | Model based on limited data may be inaccurate |
| | Lag temperature/ degree-day | Tends to reduce bias for weather-sensitive accounts | Tends to increase variability of baseline estimate. |
| | Conditional | Allows same general form and procedure to be used for weather-sensitive and non-weather-sensitive accounts, without pre-screening. Doesn't add much error for non-weather-sensitive accounts. | More complex. May give less consistent results across events for an account, if weather terms are sometimes retained and sometimes not. |
| Adjustment to precurtailment hours | Any | Simple, easy to use and understand, low cost | May be potential for gaming behavior during day-of-curtailment adjustment period |
| | | Adjusts to weather and operating conditions of curtailment day | Appropriate pre-curtailment increase in load (e.g., pre-cooling) will result in overstated baseline |
| | | Limits potential for collecting windfall credits for planned shut-downs | Pre-curtailment decrease in load in response to curtailment request (e.g., long ramp-down, canceling a shift) will result in understated baseline |
| | Additive | May adjust well for load change that is constant throughout day (e.g., industrial processes) | May not be appropriate if load changes during curtailment period (ratio adjustment may be better suited) |
| | Scalar | May adjust well for load change that is function of exogenous factor throughout day (e.g., higher levels of occupancy) | May not be appropriate if the day-to-day load variation is constant over the day (additive adjustment may be better suited) |
| | to last 2 hours before curtailment period | If load in these hours is unaffected by anticipated or initiated curtailment, provides best accuracy | If substantial curtailment is initiated in these hours, severely understates baselines |
| | to 3rd and 4th hour before curtailment period | Less potential for understated baseline due to pre-curtailment-period demand response | More variability than adjustment to last 2 hours |
| Weather-Based Adjustment | Any | Explicitly takes into account weather conditions | Adjustment may not be known to customer until after curtailment period (i.e., until after weather conditions are known for the day) |
| | | No opportunity for gaming as with adjustment to precurtailment hours | If no observations are available for extreme conditions, estimates used for adjustment may be outside range of model |
| | | | Will badly predict load reductions if the buildings are dominated by internal loads |
| | | | Less accurate than alternative adjustments or weather model for both weather-sensitive and non-weather-sensitive accounts |

Further details on the test results are provided below. Recommendations based on both the practical design considerations and the test results are then presented.

## 6.2 FINDINGS FROM TESTS ON UNCURTAILED ACCOUNTS

Test results for both weather-sensitive and non-weather-sensitive summer accounts for the best performing methods are summarized in Table 6-2. Qualitative conclusions from the test findings for both summer and nonsummer are described below.

### 6.2.1 Weather Models

#### Summer Weather-Sensitive Accounts

- For summer weather-sensitive accounts, best overall performance (low bias and low variability) is a weather model with adjustment.
- Performance is not dramatically different for different models tested, except that the model with lagged degree-day term showed higher variability than the others.
- The unadjusted lag model based on a full season of data had the lowest bias of any unadjusted method, about 0.2 percent. But with adjustment the bias was higher than with almost any other method (+ 3 to 4 percent).
- Simple models based on daily temperature or daily degree-days performed about as well as a model with hourly degree-days.
- Adjusted averages can perform nearly as well as weather models, but tend to have either worse bias or worse variability or both.
- Without adjustment, all the weather models except the full-season lag model still tend to understate the baseline, though not by as much as the unadjusted averages.
- Both bias and variability are reduced by longer input data series.

#### Summer Non-Weather-Sensitive Accounts

- Use of a "conditional" weather model (that drops weather variables for an account if they're not statistically significant or tend to have the wrong sign) doesn't increase variability compared to using an average, and appears to reduce some potential for bias.
- For high variability accounts, the lowest variability and bias is for the full season conditional weather models. With less than a full season of data, many of the adjusted averages perform about as well. However, the "high 5" (highest 5 of the last 10) has higher variability than the other averages.
- Adjustment to the hours before curtailment is important for both averages and weather models.

**Table 6-2**
**Summary of Test Results for Summer Accounts**

| Load Type | Estimation | Data Selection | Adjustment | Median Relative Hourly Error | | | Median Relative Root-Mean-Square Error | | | Extreme Relative Root-Mean-Square Error | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Weather-Sensitive Low Variability** | Weather Model | last 10 | add 1-2 | -0.2% | to | 0.7% | 5.8% | to | 7.4% | 15.2% | to | 19.2% |
| | | last 20 | add 1-2 | -0.3% | to | 1.1% | 5.3% | to | 7.5% | 15.5% | to | 22.3% |
| | | full season | add 1-2 | -0.2% | to | 2.9% | 5.1% | to | 6.9% | 14.7% | to | 19.2% |
| | Average | High 5 of 10 | add 1-2 | | 0.3% | | | 5.6% | | | 15.2% | |
| | | last 10 | add 1-2 | | -0.6% | | | 5.6% | | | 14.9% | |
| | Weather Model | last 10 | none | -2.7% | to | -1.6% | 8.1% | to | 10.7% | 28.6% | to | 30.7% |
| | | last 20 | none | -3.3% | to | -0.7% | 7.9% | to | 8.6% | 23.4% | to | 28.2% |
| | | full season | none | -2.7% | to | 0.9% | 6.6% | to | 8.2% | 24.4% | to | 26.9% |
| | Average | High 5 of 10 | none | | -4% | | | 7% | | | 25% | |
| | | High 10 of 11 | none | | -7% | | | 8% | | | 27% | |
| | | last 10 | none | | -9% | | | 11% | | | 29% | |
| | Average | High 5 of 10 | scalar THI | | 0.6% | | | 5.8% | | | 25.8% | |
| | | | add THI | | 0.4% | | | 5.9% | | | 26.0% | |
| | Average | High 5 of 10 | add 3-4 | | 0.1% | | | 7.3% | | | 22.9% | |
| | | last 10 | add 3-4 | | -1.5% | | | 7.2% | | | 23.0% | |
| | | last 20 | add 3-4 | | -1.6% | | | 6.6% | | | 24.8% | |
| **Weather-Sensitive High Variability** | Weather Model | last 10 | add 1-2 | -1.5% | to | 0.0% | 12.8% | to | 15.7% | 158.6% | to | 208.5% |
| | | last 20 | add 1-2 | -1.9% | to | 0.0% | 11.4% | to | 13.8% | 128.6% | to | 207.6% |
| | | full season | add 1-2 | -2.0% | to | 1.2% | 10.3% | to | 14.1% | 136.8% | to | 136.8% |
| | Average | High 5 of 10 | add 1-2 | | 0.6% | | | 11.9% | | | 283.4% | |
| | | last 10 | add 1-2 | | -1.1% | | | 10.3% | | | 158.6% | |
| | Weather Model | last 10 | none | -11.6% | to | -8.4% | 19.0% | to | 19.5% | 272.0% | to | 272.0% |
| | | last 20 | none | -8.7% | to | -5.4% | 21.6% | to | 23.3% | 240.5% | to | 240.5% |
| | | full season | none | -7.7% | to | -1.5% | 17.8% | to | 25.0% | 264.5% | to | 264.5% |
| | Average | High 5 of 10 | none | | -9.4% | | | 23.1% | | | 551.7% | |
| | | High 10 of 11 | none | | -21.4% | | | 27.6% | | | 331.2% | |
| | | last 10 | none | | -22.7% | | | 28.9% | | | 272.0% | |
| | Average | High 5 of 10 | scalar THI | | -3.9% | | | 18.5% | | | 546.7% | |
| | | | add THI | | -5.1% | | | 20.1% | | | 549.1% | |
| | Average | High 5 of 10 | add 3-4 | | 3.4% | | | 14.0% | | | 325.4% | |
| | | last 10 | add 3-4 | | -1.7% | | | 13.7% | | | 222.0% | |
| | | last 20 | add 3-4 | | -2.4% | | | 15.6% | | | 217.2% | |

**Table 6-2 (cont)**
**Summary of Test Results for Summer Accounts**

| Load Type | Estimation | Data Selection | Adjustment | Median Relative Hourly Error | | | Median Relative Root-Mean-Square Error | | | Extreme Relative Root-Mean-Square Error | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Nonweather-Sensitive Low Variability** | Weather Model | last 10 | add 1-2 | -0.8% | to | 0.8% | 5.2% | to | 6.4% | 21.0% | to | 45.5% |
| | | last 20 | add 1-2 | -0.6% | to | 0.1% | 5.1% | to | 5.9% | 18.6% | to | 27.1% |
| | | full season | add 1-2 | -0.7% | to | 0.3% | 5.0% | to | 5.6% | 19.3% | to | 19.3% |
| | Average | High 5 of 10 | add 1-2 | 0.1% | | | 5.0% | | | 21.4% | | |
| | | last 10 | add 1-2 | -0.7% | | | 5.3% | | | 20.2% | | |
| | Weather Model | last 10 | none | -0.8% | to | 1.1% | 9.8% | to | 10.9% | 29.5% | to | 51.1% |
| | | last 20 | none | -0.6% | to | 1.0% | 8.8% | to | 9.0% | 25.2% | to | 32.4% |
| | | full season | none | -0.8% | to | 0.4% | 7.8% | to | 8.5% | 26.2% | to | 26.2% |
| | Average | High 5 of 10 | none | 3% | | | 8% | | | 40% | | |
| | | High 10 of 11 | none | 0% | | | 8% | | | 39% | | |
| | | last 10 | none | -1.3% | | | 9.5% | | | 41.5% | | |
| | Average | High 5 of 10 | scalar THI | 3.0% | | | 9.4% | | | 40.0% | | |
| | | | add THI | 2.9% | | | 9.4% | | | 39.9% | | |
| | Average | High 5 of 10 | add 3-4 | 0.6% | | | 6.7% | | | 27.6% | | |
| | | last 10 | add 3-4 | -1.5% | | | 7.1% | | | 31.9% | | |
| | | last 20 | add 3-4 | -1.1% | | | 7.0% | | | 25.3% | | |
| **Non-Weather-Sensitive High Variability** | Weather Model | last 10 | add 1-2 | -1.3% | to | -0.6% | 20.9% | to | 23.3% | 49.4% | to | 50.7% |
| | | last 20 | add 1-2 | -1.3% | to | -0.7% | 21.7% | to | 23.2% | 50.5% | to | 50.5% |
| | | full season | add 1-2 | -0.5% | to | -0.2% | 23.1% | to | 24.1% | 88.6% | to | 88.6% |
| | Average | High 5 of 10 | add 1-2 | 2.3% | | | 22.6% | | | 58.8% | | |
| | | last 10 | add 1-2 | -1.0% | | | 20.9% | | | 49.4% | | |
| | Weather Model | last 10 | none | -5.6% | to | -4.0% | 21.9% | to | 23.0% | 56.2% | to | 62.4% |
| | | last 20 | none | -7.6% | to | -6.1% | 24.4% | to | 27.7% | 83.3% | to | 83.3% |
| | | full season | none | 2.5% | to | 3.5% | 32.9% | to | 35.1% | 260.0% | to | 260.0% |
| | Average | High 5 of 10 | none | 9.3% | | | 23.4% | | | 100.0% | | |
| | | High 10 of 11 | none | -2.7% | | | 20.7% | | | 62.3% | | |
| | | last 10 | none | -5.6% | | | 21.9% | | | 62.4% | | |
| | Average | High 5 of 10 | scalar THI | 6.6% | | | 25.5% | | | 84.4% | | |
| | | | add THI | 4.8% | | | 28.9% | | | 85.0% | | |
| | Average | High 5 of 10 | add 3-4 | 5.8% | | | 24.1% | | | 70.5% | | |
| | | last 10 | add 3-4 | -2.2% | | | 21.8% | | | 51.2% | | |
| | | last 20 | add 3-4 | -1.9% | | | 20.5% | | | 55.5% | | |

### *Nonsummer Accounts*

**Weather-Sensitive Accounts**

- Accurate weather modeling is more difficult than for summer accounts. The performance of weather models is not clearly better than that of the averages.
- For low-variability accounts, an additive adjustment with any of the averages except the "high 5 of 10", and any of the models has a median relative error close to zero.
- For high-variability accounts, the adjusted averages have median relative error close to zero, while the weather models show some negative bias. The averages also show somewhat lower variability than the weather models.

### *Nonsummer Non-Weather-sensitive*

- For nonsummer non-weather-sensitive accounts there is little difference across estimation methods and selection rules in either bias or variability.
- For both weather models and averages, additive adjustment to the two hours before curtailment tends to produce the lowest bias and variability of any of the possible adjustments.
- Unadjusted averages and weather models tend to understate load.

## *6.2.2 Adjustments*

- Some kind of adjustment to hours before curtailment helps to reduce bias and variability even with weather models.
- Additive adjustment to two hours before generally performed best of the adjustments tested, in terms of both bias and variability reduction.
- Scalar adjustment to two hours before was often as good or better, but in some cases blew up and produced much higher variability.
- PJM's THI-based adjustment generally had worse variability than the other adjustments, and was not superior in bias reduction.

## *6.2.3 Data Selection*

- Bias and variability of weather models is reduced by longer input data series, but not dramatically.
- The increased variability with shorter input series is more noticeable for conditional weather models applied to non-weather-sensitive accounts, particularly high variability accounts.

- The different averages compared performed similarly in terms of bias and variability, except for those that select a subset of days based on high load. For summer loads, the High 5 of 10 average reduces the otherwise negative bias. For summer loads using additive adjustment, High 5 of 10 gives the lowest bias of any of the averages, for both weather-sensitive and non-weather-sensitive accounts, and comparable variability. The 10 of 11 average gives some bias reduction, but not as much. For nonsummer loads, however, the High 5 of 10 average inflates an already positive bias. The other averages perform better, and roughly comparably to each other, in terms of both bias and variability, for both weather-sensitive and non-weather-sensitive accounts.

## 6.3 Findings from Tests on Curtailed Accounts

A key finding from testing curtailed accounts is the practical challenge of using adjustments to the hours just before curtailment. For uncurtailed accounts, adjustment to the last two pre-curtailment hours reduced both bias and variability for every estimation method and selection rule, except in some cases for the high 5 of 10 average. In most cases, this adjustment had lowest bias and lowest variability. This result makes senses because the hours closest to the curtailment period are likely to be most indicative of both weather and operating conditions for that day, in the absence of curtailment.

However, many of the actual curtailed accounts indicated curtailment, often to close to zero load, by the start of the two hours prior to the official start of the curtailment period. These are not cases of plant shut-down, since the earlier hours are at typical high loads. Rather, the operations appeared to be curtailed rapidly after a curtailment notification. In these cases, the adjustment to the two hours before the curtailment period produced very low, sometimes negative, baselines, with severe understatement of curtailment amounts.

For this reason, the additive adjustment is not a practical choice for accounts that are likely to implement load reductions in this way. As discussed in Section 3, this adjustment is also a concern because of the possibility that customers will deliberately boost load in the hours prior to curtailment to produce an artificially high baseline. Pre-cooling the building, without intent to "game" the system, can produce a similar distortion.

Thus, while the additive adjustment to the last two hours is very effective for accounts that will not change operations substantially outside of the curtailment period, this method has problems for other accounts. A different method is needed for accounts whose operations outside of curtailment periods is substantially affected by a notification or anticipation of a curtailment event.

Alternatives include the following:

1. Adjust to the third and fourth hours prior to curtailment. This approach reduces, but does not eliminate, the problem related to reductions before the start of the formal curtailment period, such as sending a shift home. The adjustment to earlier hours also reduces the

potential for manipulation of the baseline, if notification is typically less than three hours from curtailment start. On the other hand, earlier hours are less indicative of what load would have been for the rest of the day absent curtailment. Test results for averages on uncurtailed accounts found that the adjustment to hours –3 and –4 adds some negative bias, and increases the variability compared to the adjustment to hours –1 and –2. For weather-sensitive accounts, if the adjustment to hours –3 and –4 is used, regression models will mostly eliminate the bias, but will also add additional variability unless at least a 20-day period is used for the regression.

2. Use a full-season regression model without adjustment. This method is essentially unbiased and has minimal variability for weather-sensitive accounts. The practical difficulty is waiting until a full season of data are available, or relying on a previous season. In addition, if operating practices have changed since the previous season, or change within the current season, the full-season model may not be a good representation of load on the curtailment day.
3. Use the "High 5 of 10" average with THI adjustment. This adjustment is not dependent on actual behavior in the hours immediately before curtailment. As a result, it is subject neither to manipulation by the customer nor to severe distortion due to curtailment in advance of the required time, nor to inflation due to pre-cooling. This combination has variability no worse than that of unadjusted averages, and for low-variability weather-sensitive accounts, it is essentially unbiased. For non-weather-sensitive accounts, it produces some positive bias and increases variability compared to unadjusted averages.

   The THI adjustment for this analysis was based on a full season of load and weather data from the season in which the curtailment occurred. Actual implementation of the THI adjustment by PJM requires the THI regressionwould be based on a full season of data from the previous year or a minimum of a month of data from the current season. Practical considerations are the following.

   - It either requires load history from the prior year or must rely on shorter data spans from the ongoing season.
   - Using previous year data, if the load pattern has changed it can produce erroneous results. Using limited data from earlier in the summer, regression results may be forced to make predictions outside of the range of the data.
   - It involves an adjustment that is difficult to explain succinctly and may be difficult to understand.

## 6.4 Recommendations

The choice of a baseline method needs to balance a number of practical considerations as well as prediction accuracy. These considerations, listed at the beginning of this section, are:

- simplicity of calculation

- minimizing burden on participants and operators, in terms of costs, ease of understanding, and ease of operation
- limiting the potential for gaming
- limiting the potential for distortion if load is curtailed early
- ability to know the baseline immediately after a curtailment
- ability to know the baseline before making a curtailment decision
- minimizing method bias: systematic tendency to over- or under-state
- minimizing method variability: tendency to wide swings in estimates.

Different methods are appropriate for different types of accounts, and according to the importance assigned to each of the above considerations.

### *Adjustments*

For almost all basic methods, an additive adjustment to the two hours prior to curtailment can reduce both bias and variability. The problems with this approach are the potential for gaming, and the potential for understatement of load reduction if a customer curtails prior to the formal curtailment period. The additive adjustment to the last 2 hours should be considered as an option for accounts that are considered unlikely to be subject to either of these distortions.

For weather-sensitive accounts, alternatives to this adjustment include:

- PJM's High-five average with THI adjustment
- Adjust to the third and fourth hour prior to curtailment.
- For summer loads, weather model using a full season of data, with no adjustment.

In general, scalar and additive adjustments perform similarly well, except that scalar adjustments sometimes blow up resulting in higher variability. If scalar adjustments are used, some procedure to avoid extreme adjustment ratios should be incorporated. On the other hand, additive adjustments can produce negative baselines; the baseline should be truncated at zero if additive adjustments are used.

### *Weather Modeling*

If more complex methods are acceptable, weather models are recommended for summer weather-sensitive accounts. Weather models may also be useful for winter weather-sensitive accounts, but are not as clearly superior to simple averages, particularly if adjustments to the current day are also used.

No single model structure offers a clear preference for all situations. Models relying on daily temperature or degree-day variables appear to perform as well as those using hourly weather inputs. However, these models are likely to behave less well on days when there are unusual changes in the weather, or in regions where such changes are likely. Likewise, models incorporating humidity and/or lag degree-day variables did not perform better across the cases examined in this study, but these terms may be important for some regions and customer types.

If weather models are used, screening criteria should be included to eliminate weather terms that are not statistically significant or physically meaningful for a particular account. This screening can be done once, at the beginning of the curtailment season or at initial enrollment in a program, and may be based on a full season of data if available. Without such screening, high variability can result for some accounts. Use of such screening procedures can allow a single general procedure, including this screening, to be applied to all accounts, without pre-classifying the accounts by weather sensitivity. Even if the weather modeling is restricted to accounts believed to be weather-sensitive, use of the screening procedure can still eliminate extreme anomalous results.

Models based on a full season of data tend to be more reliable than those based on shorter periods. However, waiting until the season has ended to determine baselines is impractical for most programs. Models based on the previous 20 business days perform slightly better than those based on the previous 10. In most cases, the difference is not dramatic, but reliance on only 10 days of data will increase the potential for anomalous events to skew results.

An alternative to using an additive or scalar adjustment after applying the weather model is to include load at hours just before curtailment as predictors in the model. This approach was not tested in the study but should be considered.

### *Averages*

If simpler models are a priority, hourly averages with adjustment to the hours prior to curtailment can be used with good results even for weather-sensitive accounts. For non-weather-sensitive accounts, averaging methods make more sense than weather models. However, adjustment to hours prior to curtailment, subject to the concerns discussed above, can improve method performance for both types of account. With this adjustment, averages based on different selection rules (last 10, last 11, highest 10 of last 11, highest 5 of last 10, or last 20) tend to perform similarly. For summer loads, the High 5 of 10 average with additive adjustment performed somewhat better than averages using other selection rules. Averages based on a full season can do somewhat better, if this much delay in results is acceptable.

### *Proposed Approaches by Load Type*

#### Offering Options

A general recommendation is that baseline calculation protocols should provide for alternatives based on customer load types and operating practices. One way to simplify the provision of options is to establish a default method, and allow certain deviations.

The basis for the selection of method should be not just the business type, but also the load patterns evident in the data as well as the customer's description of operating practices. Thus, for example, a customer who indicates a desire to be able to cancel a shift in advance of the control period should have access to a baseline calculation method that is not distorted by this practice.

At the same time, the program operator should have some discretion to bar customers from using an approach that they appear to have manipulated in the past. Thus, if there is evidence that a particular customer tends to inflate load after notification, beyond what would reasonably be expected for pre-cooling, that customer might not be able to use a method that includes adjustment to the two pre-curtailment hours.

#### A Practical Default Baseline Calculation Method

A method that generally works well for a range of load types is the simple average of the last 10 days, with additive adjustment to the two hours prior to the curtailment period. This method can be recommended for both weather-sensitive and non-weather-sensitive accounts, with both low and high variability, for summer and nonsummer curtailments.

This method is not recommended for accounts that tend to begin load curtailment substantially in advance of the formal curtailment period. It is also not recommended for situations where the potential for gaming is a strong concern, whether across the program or for particular customers.

#### Alternatives for Summer Weather-Sensitive Accounts

For summer programs, practical alternatives for weather-sensitive accounts include the following:

- unadjusted weather models. Longer input time periods are preferable, particularly for high-variability accounts.
- the High 5 of 10 average with THI adjustment.

Simpler methods with less desirable but potentially acceptable performance include

- Unadjusted averages, particularly the High 5 of 10
- Averages or weather models adjusted to the third and fourth hour before curtailment.

#### Alternatives for Summer Non-Weather-Sensitive Accounts

For non-weather-sensitive summer accounts, the unadjusted High 10 of 11 average performs nearly as well as the recommended default, particularly for low-variability accounts. Next best is the simple average of the last 10 days with additive adjustment to the third and fourth hours before curtailment.

For low-variability accounts, unadjusted weather models, with weather terms retained only if indicated by the data, actually perform slightly better than the recommended default. However, unlike the case for weather-sensitive accounts, these models perform better if based on shorter periods of data. For high-variability accounts, unadjusted weather models tend to be worse than the unadjusted high 10 of 11 average.

### Alternatives for Nonsummer Accounts

For nonsummer accounts, modeling is more challenging and there are fewer alternatives. For weather-sensitive accounts, the High 5 of 10 average with THI adjustment can be used. For low-variability accounts, the unadjusted High 5 of 10 average appears to perform slightly better, but for high-variability accounts it is worse.

For non-weather-sensitive nonsummer accounts, the unadjusted High 10 of 11 appears to be the best alternative. Any of the averages with additive adjustment to the third and fourth hour before curtailment perform not quite as well.

### Summary of Recommended Methods and Alternatives

The recommended methods, alternatives, and operational considerations for different load types are summarized in Table 6-3.

**Table 6-3**
**Recommended Methods and Alternatives**

| Season | Weather Sensitivity | Variability | Recommended Default Estimation | Recommended Default Data Selection | Recommended Default Adjustment | Recommended Alternatives Estimation | Recommended Alternatives Data Selection | Recommended Alternatives Adjustment |
|---|---|---|---|---|---|---|---|---|
| Summer | Weather-Sensitive | Low | Average | last 10 | add 1-2 | weather models | any | none |
| | | | | | | Average | High 5 | THI |
| Summer | Weather-Sensitive | High | Average | last 10 | add 1-2 | weather models | longer is better | none |
| | | | | | | Average | High 5 of 10 | THI |
| Summer | Non-Weather-Sensitive | Low | Average | last 10 | add 1-2 | weather models | shorter is better | none |
| | | | | | | Average | High 10 of 11 | none |
| | | | | | | Average | last 10 | add 3-4 |
| Summer | Non-Weather-Sensitive | High | Average | last 10 | add 1-2 | Average | High 10 of 11 | none |
| | | | | | | Average | last 10 | add 3-4 |
| Nonsummer | Weather-Sensitive | Low | Average | last 10 | add 1-2 | Average | High 5 of 10 | none |
| Nonsummer | Weather-Sensitive | High | Average | last 10 | add 1-2 | Average | High 5 of 10 | THI |
| Nonsummer | Weather-Sensitive | Low | Average | last 10 | add 1-2 | Average | High 10 of 11 | none |
| | | | | | | Average | last 10 | add 3-4 |
| Nonsummer | Weather-Sensitive | High | Average | last 10 | add 1-2 | Average | High 10 of 11 | none |
| | | | | | | Average | last 10 | add 3-4 |