Baselining the ARENA-AEMO Demand Response RERT Trial

SEPTEMBER 2019

Oakley Greenwood
Developed with Oakley Greenwood

Australian Government
Australian Renewable Energy Agency

ARENA

DISCLAIMER

DOCUMENT INFORMATION
Project ARENA-AEMO Demand Response RERT Trial
Client ARENA
Report prepared by Lance Hoch (lhoch@oakleygreenwood.com.au)
Date September 2019

CONTENTS

1. EXECUTIVE SUMMARY

In 2017, ARENA and AEMO entered into a Memorandum of Understanding to jointly develop 'proof of concept' projects that support the integration of renewable energy into the energy market, while maintaining system reliability and security. As part of this initiative, a three-year Demand Response (DR) Short Notice Reliability and Emergency Reserve Trader (SN RERT) trial (RERT Trial) was developed to provide evidence to inform DR's role in maintaining grid security and reliability.

A key characteristic of DR, as compared to energy consumption, is that it cannot be measured directly. It is estimated by comparing actual consumption with a prediction of what would have occurred if the request for DR had not been made. There are several approaches for generating this prediction, of which baselining – using the history of the site's demand – is the most common.

This report analyses the relative accuracy, bias and precision of the '10 of 10' baselining methodology used to predict the actual metered load for participants in the RERT Trial. Put simply, the '10 of 10' methodology uses the consumption of the 10 most recent qualifying days to construct a baseline.

The results of this analysis suggest that the '10 of 10' baseline methodology currently used in the RERT may be adequate for certain types of loads, particularly those of larger commercial and industrial customers whose energy consumption is relatively similar from day to day and not particularly weather sensitive.

Where the load shape is not relatively consistent from day to day, the '10 of 10' methodology can result in the baseline not being an accurate estimate. Considering this, the '10 of 10' methodology might not be appropriate for the following types of loads:

- **Highly weather-sensitive loads** – This was primarily an issue for residential facilities, but also for some smaller commercial facilities where weather (and particularly ambient temperature) has a material impact on total energy demand;
- **Loads influenced by rooftop PV generation** – This was only cited as an issue for residential facilities, though it applies in principle to commercial and industrial facilities where the PV generation capacity is material compared to the load providing DR;
- **Loads that vary from day to day, but in a consistent pattern** – For example, where the facility has a different level or schedule of operation on specific days of the week (this was primarily cited as an issue for commercial and industrial facilities); and/or
- **Highly intermittent loads** – For example, where the facility or specific load providing DR is driven by internal activity factors that are not related to external variables such as weather or day.

Other approaches that may offer better alternatives for these types of loads include anchoring or the use of control groups. Anchoring assesses the shape of consumption of the facility on days of like temperature in the past and the pre- and post-period consumption of the facility on the event day to construct the baseline.

A control group is a group of customers whose consumption on event days can be assumed (or has been shown) to be similar to that of the customers providing DR. The difference between consumption on the day of the DR event of the control group and the DR customers is taken to represent the amount of DR delivered.

2. BACKGROUND

2.1 The ARENA-AEMO Demand Response RERT Trial

In 2017, ARENA and AEMO entered into a Memorandum of Understanding to jointly develop ‘proof of concept’ projects that support the integration of renewable energy into the energy market. As part of this initiative, a three year DR Short Notice RERT Trial was developed to:

- Evaluate the performance of various demand response (DR) resources in electricity supply contingency events.
- Improve the commercial and technical readiness of innovative DR approaches, such as engagement with mass market customers through behavioural and technology-enabled responses.
- Provide an evidence base to inform the design of a new market, or other mechanisms, for provision of demand response to assist with grid reliability and security.

In total, $35.7 million of funding was provided, with ARENA providing up to $28.55 million of funding, and the NSW Government providing $7.18 million for NSW projects (funded in a 50/50 split with ARENA). Proponents within the trial are required to sign on to the AEMO Short Notice RERT Panel and make their DR capacity available if and when requested, at 2017 prices. Proponents receive the ARENA grant in the form of availability payments that are provided after semi-annual test dispatches confirm the proponent’s ability to deliver the amount of DR they have been contracted for. Proponents also receive usage payments from AEMO that are capped at $1,000/MWh for demand response activated under the RERT process.

2.2 The role of baselines in DR programs

DR is defined as the reduction in demand that occurs at a connection point as a result of a specific request to the end user from another party. A key characteristic of DR – as compared to energy consumption – is that it cannot be measured directly. It is estimated by comparing actual consumption with a prediction of what would have occurred if the request for DR had not been made.

There are several approaches for estimating DR[1], of which baselining – using the history of the site’s demand – is the most common. The level and pattern of consumption on other days (when no request for DR was made) is taken as the baseline for the customer or group of customers providing DR. The difference between the baseline and the metered consumption that occurs in response to the DR request is deemed to be the amount of DR delivered.

There are a number of variables to consider when developing a baseline:

- How the ‘history’ used in the approach is defined. This includes:
  - What constitutes a ‘qualifying day’ in the baseline calculation[2] – for example, should the approach consider the same type of day (weekday versus weekend) as the DR day, or whether the historical days should be similar to the DR day in other respects, such as weather conditions.
  - How many qualifying days should be considered when developing the baseline.
  - How far back in time you can go in selecting qualifying days.
- How the approach adjusts the baseline for consumption prior to the start and immediately after the end of the DR period.

1 Other common approaches for quantifying the amount DR provided during a DR event include anchoring and the use of a control group.

2 ‘Qualifying days’ are defined in the RERT Trial contract (and in AEMO’s agreements for DR in all other parts of the RERT) as “calendar weekdays… which are not public holidays (in that location) and on which demand response events have not been called for the NMI.”

3. BASELINING IN THE RERT

The baseline methodology selected by AEMO for use in the RERT Trial – and for DR contracted in other parts of the RERT – was developed by the California Independent System Operator (CAISO) and is known as the '10 of 10' baseline.

This approach uses the consumption of the 10 most recent qualifying days to construct the baseline. Qualifying days are defined as being either the 10 most recent weekdays, if the DR event takes place on a weekday, or the 10 most recent weekend days, if the DR event occurs on a weekend.[3] The CAISO '10 of 10' methodology was chosen after an examination of several baselining approaches being used internationally at the time[4] and determined that, on average, it provided a more accurate baseline and entailed lower bias[5] than the other approaches examined.

It should be noted that baselining can be used to measure the DR delivered by a single customer or a group of customers in aggregate. In the RERT Trial, AEMO applies the baseline to the aggregated customer consumption of the trial proponent ('portfolio level').[6]

3.1 Perceived shortcomings of the '10 of 10' methodology

The CAISO '10 of 10' baseline load profile is derived from the 10 preceding qualifying days, adjusted for actual consumption on the day of the demand response event. This works best for facilities with load profiles that are quite consistent from day to day, such as the large industrial and some commercial loads that have been the traditional sources of DR in Australia and the USA.

However, where the load shape is not relatively consistent from day to day – and particularly where the load shape on an event day is different to the average load shape – the CAISO '10 of 10' method can result in the baseline not being an accurate estimate of what the consumption would have been on a DR event day in the absence of DR being provided.

The experience of several of the proponents in the first year of the RERT Trial suggested that the '10 of 10' methodology might not be appropriate for the following types of loads:

- **Highly weather-sensitive loads** – This was primarily an issue for residential facilities, but also for some smaller commercial facilities where weather (and particularly ambient temperature) has a material impact on total energy demand.
- **Loads influenced by rooftop PV generation** – This was only cited as an issue for residential facilities, though it applies in principle to commercial and industrial facilities where the PV generation capacity is material as compared to the load providing DR.
- **Loads that vary from day to day, but in a consistent pattern** – For example, where the facility has a different level or schedule of operation on specific days of the week (this was primarily cited as an issue for commercial and industrial facilities).
- **Highly intermittent loads** – For example, where the facility or specific load providing DR is driven by internal activity factors that are not related to external variables such as weather or day type. This was only cited as an issue for relatively large industrial loads but could also be relevant to individual residential and small non-residential facilities, for example a small industrial facility with intermittent but material machine loads[7].

3 The RERT baseline approach does not currently make explicit provision for RERT events on weekend days.

4 Development of Demand Response Mechanism: Baseline Consumption Methodology – Phases 1 and 2, published July and October 2013 by AEMO; see www.aemo.com.au.

5 AEMO, *DRM Detailed Design*, 2013.

6 Schedule 2 to the RERT Panel Agreement states: "The aggregated electricity demand of all *NMI*s and *datastreams* in the list provided by the *Reserve Provider* to *AEMO* after *activation* will be used to calculate the baseline and the amount of *reserve activated*. Baselines and *reserve activated* will not be calculated for individual *NMI*s and *datastreams*"

7 An important difference at the portfolio level is that this effect in a small customer facility would be unlikely to change the portfolio outcome, whereas such an occurrence in a large facility might do so.

3.2 The approach taken in this assessment

Modelling was conducted to assess the relative accuracy, bias and precision of the '10 of 10' and alternative baseline methodologies in predicting actual metered load for participants in the RERT Trial. The modelling was conducted using days and times similar to those on which the RERT would be most likely to be called (i.e. at times of very high demand). The approach taken in the modelling included the following steps:

1. Developing simulated event days

2016-17 half-hourly metering data was obtained for all sites that had signed up to participate in the RERT Trial for the 2017-18 summer.

The 30 days of highest demand in the 2016-17 year in each of the three jurisdictions that participated in the Trial were identified. All of those that occurred on weekdays from 1 December 2016 through 31 March 2017 were selected as potential DR event days.[8] The assumed 4-hour potential RERT event period on each day was defined as extending from two hours before until two hours after the time of peak demand on that day. While a RERT activation could occur at any time, it is most likely to occur when demand is at its highest, as it is at those times when the balance between the available supply and demand is likely to be tightest, which in most places is on hot summer weekdays – particularly when several hot days occur in a row.[9]

The strength of this approach is that the data for the event day was what the participants actually did consume in the absence of a call for demand response, and this could be compared to the various baselines that had been constructed to determine how well that baseline predicted consumption on an event day.

2. Testing the suitability of the '10 of 10' approach by customer class and jurisdiction

For each customer class within each jurisdiction, the '10 of 10' approach was used to develop baselines for each national metering identifier (NMI) for each of the 30 simulated event days.

The accuracy, bias and precision of the '10 of 10' baselines in predicting the actual metered load of each site across the 30 days was then assessed using the statistics discussed in Section 3.4.

3. Testing the ability of alternative approaches to provide better baselines

Where the '10 of 10' method did not provide sufficient accuracy for a customer class within a jurisdiction, alternative baselines (see Section 3.3) were developed.

The alternative baselines were then subjected to the same assessment of accuracy, bias and precision in predicting the actual metered load of each site across the 30 simulated DR event days.

4. Comparing results and identifying the preferred baseline approach or additional analysis needed

The results of the '10 of 10' and alternative baseline approaches were compared for each load type.[10] Baselines that achieved higher levels of accuracy and precision and lower levels of bias were preferred. These results are discussed in the *Findings* section below.

---

8 Most of the 30 highest demand days in each of the jurisdictions did occur within this timeframe. Where one of the highest 30 demand days occurred outside those dates the next highest demand day within those dates was selected.

As a further check, baselines were developed and analysed using the approach described here for every day from 1 December 2016 through 31 March 2017. This report presents the results for the 30 highest demand days, as they are more likely to be similar to conditions on days when the RERT is activated.

9 It is very unlikely that more than a few of the end users participating in the ARENA program would have been asked to provide demand response during the 2016-17 summer because the RERT program was not operating in that year, and most of the participants in the ARENA-AEMO RERT DR Trial were not actively involved in DR programs prior to their involvement in the Trial.

10 NMI metering data was analysed to establish the customer class and load type of the NMIs within the proponents' portfolios, as a significant proportion of the classifications originally provided by the proponents proved unreliable.

3.3 Alternative baseline methodologies tested

Alternative baseline methodologies were selected based on the work of the CAISO Baseline Adequacy Working Group (BAWG).[11] In 2017-18, the BAWG undertook a major review of the applicability and accuracy of the '10 of 10' baselining approach to a wider variety of load types, with significant attention to residential DR.

Based on the BAWG's work, the following alternative baseline approaches were tested for three of the four load types for which improvements to the '10 of 10' were being assessed:[12]

- '10 of 10' (current RERT trial method, including its 20% adjustment factor)
- Like days in terms of maximum temperature, including a larger (40%) adjustment factor[13]
- Like days in terms of average temperature, including a larger (40%) adjustment factor
- Like day of the week and maximum temperature, including a larger (40%) adjustment factor
- Like day of the week and average temperature, including a larger (40%) adjustment factor.

3.4 Key evaluative measures used

The comparative analysis of baseline methodologies sought to identify the approach that would provide the most suitable baseline in terms of three evaluation measures[14], each of which was calculated at the relevant portfolio level:

- **Accuracy** – The degree to which the baseline is able to accurately predict energy demand.[15] Accuracy is considered to be the most important of the three evaluative measures.[16]
- **Bias** – The degree to which the baseline method tends to over- or under-predict the actual metered load of the portfolio.[17] Most programs seek baselines with zero bias; however, baselines characterised by consistent but minor under- or over-estimates are acceptable as any residual error will be known and an adjustment factor can be considered.
- **Precision** – Precision refers to how reliable the result is with repeated trials.[18] A precise outcome will show the same (or a very similar) result each time the method is employed. Where the results are precise, they can be adjusted for accuracy.

---

11 The BAWG's alternative baselines were selected for assessment because they were developed with the specific purpose of improving the applicability and accuracy of the '10 of 10' approach, which had also been developed by CAISO.

12 Highly intermittent loads were not tested using any of these alternatives because these loads are driven by internal activity factors that are not related to external variables such as weather or day type. As such, seeking correlations with weather or day of the week would be unlikely to be of value. An alternative approach for these types of loads is discussed in the following sub-section.

13 The adjustment factor places a limit on the difference between the average consumption prior to (and, where a post-period adjustment factor is used, the average consumption after) the DR activation period and the consumption at that same time on the event day. The adjustment factor is used to raise (or lower) the baseline to reflect the consumption conditions of the event day. A larger adjustment factor may be appropriate where the drivers of consumption result in a marked change in consumption levels on an event day.

14 The statistical tests employed in this study were the three key parameters used for comparing potential baselining methods in a DNV-KEMA study entitled *Development of Demand Response Mechanism, Baseline Consumption Methodology – Phase 2 Results Final Report* that was commissioned by AEMO.

15 Accuracy was measured by the Relative Root Mean Square of the Errors (RRMSE).

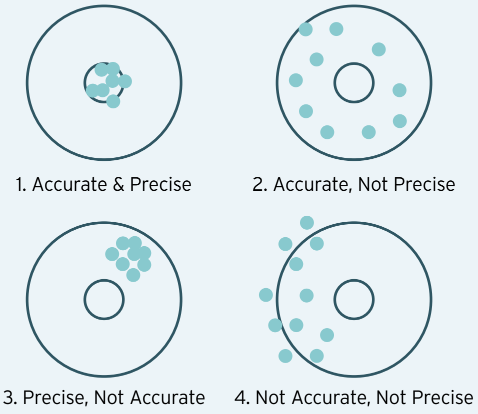
16 As stated by DNV-KEMA in their report cited in the footnote above, "The RRMSE combines the systemic errors measured by the ARE (bias) and the variability of errors captured by the RER (precision)." For this reason, the RRMSE was used as the primary evaluative measure in this study.

17 Bias was measured by the Average Relative Error (ARE).

18 Precision was measured by the Relative Error Ratio (RER).

Figure 1 illustrates the difference between accuracy and precision.

Figure 1: Accuracy and precision illustrated



4. FINDINGS

4.1 Sample size and coverage

Table 1 shows the number of sites that were included in the analysis by customer class within each of the jurisdictions. Table 2 shows the number of sites within each customer class for which alternative baseline methods based on load type were analysed.

Table 1: Number of NMIs analysed by jurisdiction and customer class[19]

	VIC	NSW	Total
Residential	6,448	0	6,448
Commercial	179	49	228
Industrial	11	26	37

Table 2: Number of NMIs analysed by customer class and load type

	'10 of 10'	Weather-sensitive	PV-influenced	Consistently variable	Intermittent	Total
Residential	6,448	6,448	1,064	0	0	6,448
Commercial	228		NA	Not determined[20]	NA	228
Industrial	26	26	NA	NA	1	26

The sample sizes for several of the load types are very small, and there were no examples of some of the load types in some of the customer classes. This limits the conclusiveness of some of the results discussed below.

19 At the time the sample of NMIs was drawn for this study, no metering information was available for any of the SA portfolios.

20 The portfolio of one NSW proponent was characterised by consistently variable loads, with Thursday evenings showing a consistently higher level of consumption that other evenings. The precise number of NMIs within this portfolio exhibiting this pattern was not determined.

How to read the RRMSE vs ARE charts

Each dot in each chart represents the accuracy and bias of the aggregate baseline for the relevant portfolio of NMIs on a specific simulated event day. Each chart shows the results for each of the simulated event days for each of the baseline methods assessed.[21]

The relative root mean square of the errors (RRMSE) on the vertical axis measures how accurately the adjusted baseline represents the load profile during the 4-hour event period.[22] The RRMSE is expressed as the ratio of the root mean square error to the average load, and is often shown as a decimal or a fraction. An RRMSE of 10 per cent or less is generally considered to be 'good', and an RRMSE between 10 and 20 per cent is considered to provide 'acceptable' accuracy.[23]

The average relative errors (ARE) score on the horizontal axis measures the bias of the baseline profile. If the ARE is negative, the adjusted baseline is underestimating the actual load on the day and therefore is a more conservative estimate of any DR. Similarly, if the ARE is positive, the baseline is overestimating the actual load on the day and will also overestimate the DR supplied.

The RER (Relative Error Ratio) was also calculated but for display purposes only the RRMSE and the ARE have been plotted.

Because the RRMSE measures accuracy it is generally used as the primary measure for selecting a baseline method.[24] Where the RRMSE for two or more baseline methods is similar, the ARE can be used as a secondary selection metric; that is, a method with less bias would be preferred. The perfect baseline method would have all of its results with an RRMSE of 10 per cent or less and zero bias, or a symmetrical distribution of ARE scores. This latter outcome would result in the plotted results forming a perfect V shape.

4.2 Industrial loads

A total of 37 NMIs with industrial loads were analysed, 11 in Victoria and 26 in NSW. The results in the two states differed considerably, as shown in Figures 2 and 3.

- The '10 of 10' method produced quite accurate baselines for the Victorian industrial NMIs.
  - In 83 per cent of the simulated event days it produced a baseline whose accuracy was +/- 10 per cent of the actual aggregate consumption of these NMIs. In 97 per cent of the simulated event days it produced a baseline of acceptable accuracy (i.e. + / – 20 per cent).[25]
  - The use of weather variables or like-day of the week with 40 per cent adjustment factor cap did not result in an improvement in the accuracy of the aggregate baseline for these NMIs.
- The accuracy of the '10 of 10' baseline was much lower for the NSW industrial NMIs.
  - The '10 of 10' approach produced a good level of accuracy (i.e. +/- 10 per cent of actual consumption) for only 11 per cent of the simulated event days for this group of NMIs, and an acceptable level of accuracy in only 29 per cent of the simulated event days. However, this was better than the accuracy produced by any of the other baseline approaches tested.
  - Examination of the sample of NSW industrial NMIs revealed that six of the 36 were relatively large (ranging in average consumption from about 5 MW to about 20 MW) and exhibited a high level of load variability on a half-hourly basis (ranging from 20 per cent to 50 per cent from one half-hour to the next). The '10 of 10' baseline approach is unlikely to provide accurate results for such loads unless their half-hourly variations are similar from day to day.

21 Although 30 simulated event days were tested with each of the baseline approaches, the graphs may contain fewer data points. This is because, for display purposes, the graphs were sized to exclude baseline results with RRMSE scores greater than 0.5.

22 It should be noted that the RRMSE only measures the similarity of the baseline to the load on the DR event day within the DR period. It does not reflect similarity or otherwise outside those hours.

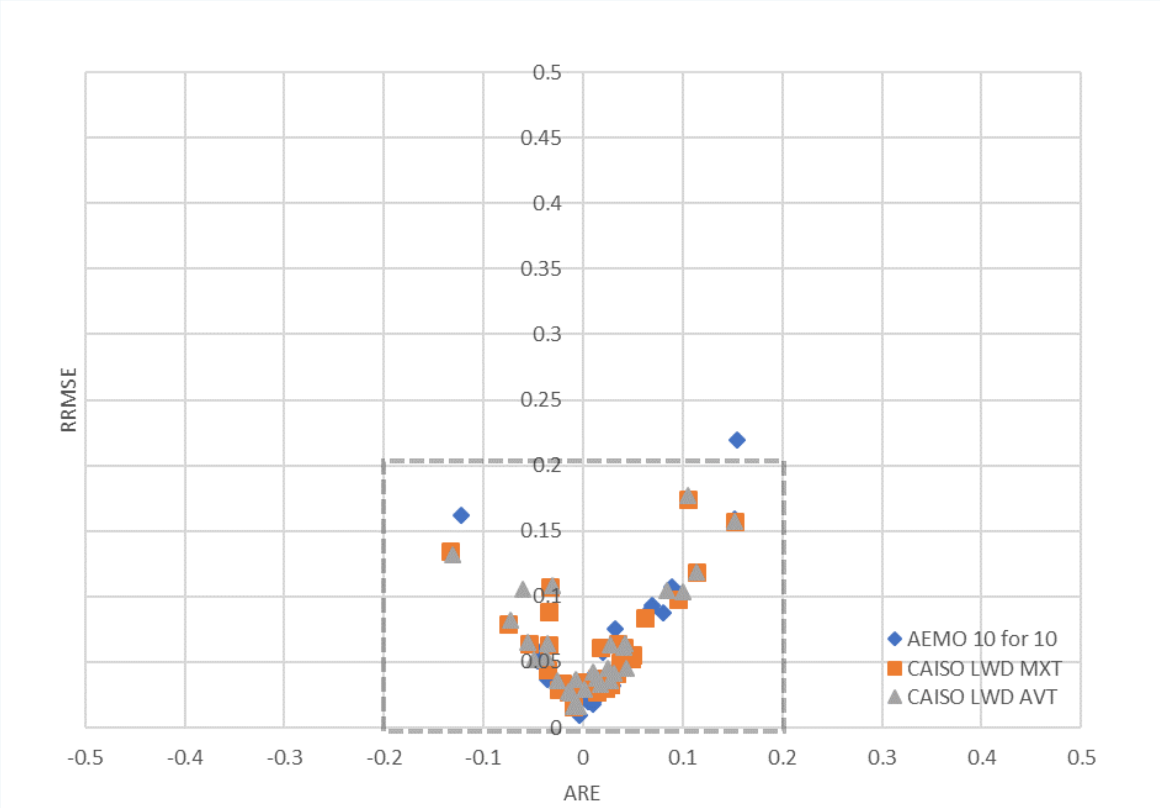
23 Several groups including AEMO, PJM and KEMA have used the 10 per cent and 20 per cent thresholds as the definition of 'good' and 'acceptable' levels of accuracy.

24 As noted by DNV-KEMA in their 2011 report *PJM Empirical Analysis of Demand Response Baseline Methods* for the PJM Markets Implementation Committee, "the RRMSE combines the systematic errors measured by the bias metric (the baseline's average relative error) and the variability of errors captured by the variability metric (relative error ratio)."

25 Establishing an appropriate confidence level – that is, setting the percentage of events in which the baseline must produce 'good' or 'acceptable' accuracy in order for that baseline approach itself to be deemed suitable for use – is essentially a policy decision. Confidence levels of 90 per cent and 95 per cent are commonly used.

Figures 2 and 3 illustrate the differences in the accuracy and bias of the various baseline approaches as applied to the industrial NMIs in Victoria and NSW.

Figure 2: Scatterplot results of accuracy and bias for VIC industrial NMIs



Note: LDW = Like weather day
MXT = Maximum daily temperature
AVT = Average daily temperature

Figure 3: Scatterplot results of accuracy and bias for NSW industrial NMIs

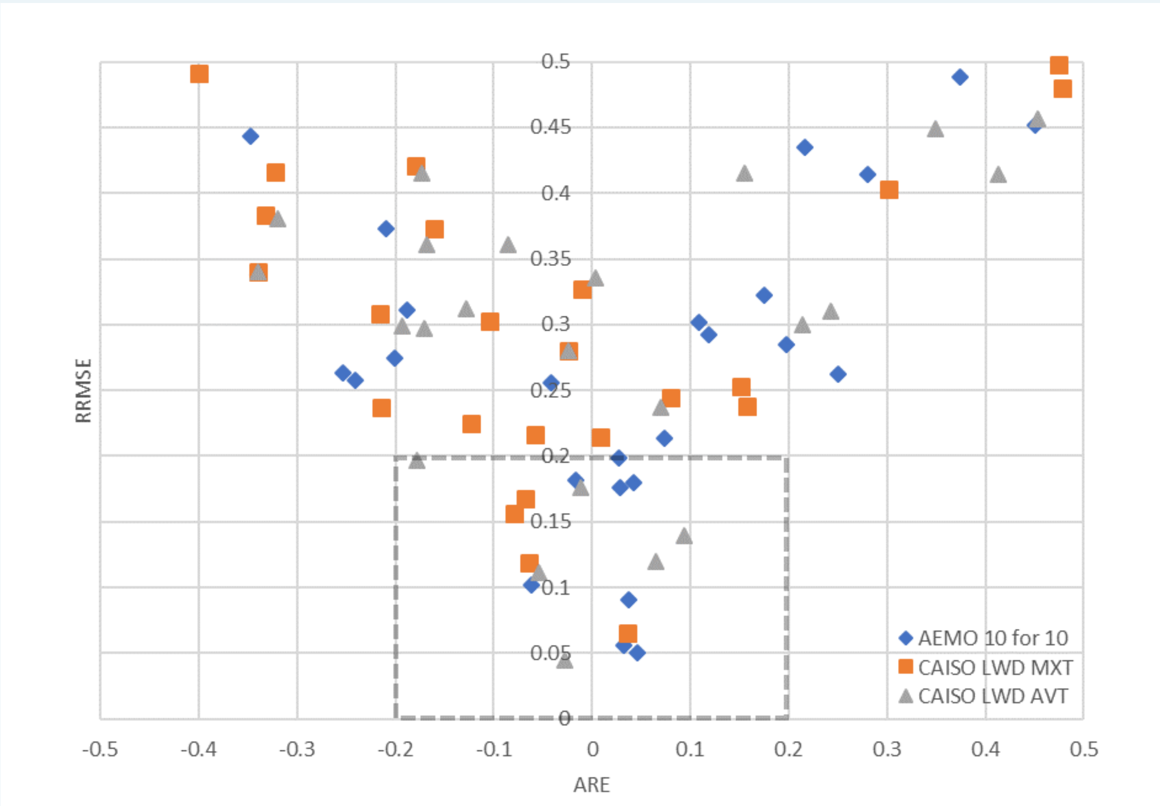


Table 3 summarises the results shown in Figures 2 and 3. The accuracy of all of the baseline approaches modelled was significantly higher in Victoria compared to NSW. None of the baseline approaches modelled in NSW performed well, possibly due to the high level of load variability within this group. In Victoria, the '10 of 10' method delivered a comparable level of accuracy to the alternative methods tested.

Table 3: Percentage of simulated event days for industrial loads in Victoria and NSW with 'good' and 'acceptable' accuracy using different baseline methods

Jurisdiction & Baseline method	Good accuracy (RRMSE < 10%)	Acceptable accuracy (RRMSE < 20%)
VIC industrial NMIs		
'10 of 10'	83%	97%
Maximum temperature	83%	100%
Average temperature	72%	100%
Day of week & maximum temperature	83%	97%
Day of week & average temperature	83%	97%
NSW industrial NMIs		
'10 of 10'	11%	29%
Maximum temperature	4%	14%
Average temperature	4%	21%
Day of week & maximum temperature	0%	11%
Day of week & average temperature	0%	11%

What defines good and acceptable accuracy?

The definition of what constitutes a 'good' or an 'acceptable' level of accuracy is essentially a policy decision. An RRMSE score of 10 per cent and 20 per cent have been used by AEMO, PJM and other organisations as thresholds of good and acceptable accuracy.

Tables 3 through 7 show the proportion of simulated event days on which the different baseline approaches achieved those levels of accuracy. The choice of the proportion of events in which a baseline approach should achieve a good or acceptable level of accuracy in order to be used is also a policy decision.

In the case of baseline accuracy in the RERT, the use of a baseline with a lower frequency of achieving good and acceptable accuracy will likely increase the total volume of DR that is available for use in the RERT, but will also;

- increase the potential for error in managing RERT events, leading to either more reserves needing to be maintained, increased potential for either load shedding to eventuate, or generators to be backed down when they may not have needed to be; and
- increase the risk that the amount of DR paid for by the market is not correct, noting that the direction of the error (over- or under-payment) will determine whether the overpayment benefits DR providers at the expense of consumers or vice versa.

Requiring a higher level of frequency of achieving good or acceptable accuracy, on the other hand, would reduce the risks of errors and incorrect payments noted above but will either:

- require the use of more and possibly more complex baselines to fit the drivers of consumption in different types of loads, particularly where there is a desire to increase the available volume of DR for use in the RERT and/or the ability of different types of customers to participate in the provision of DR for the RERT, thereby increasing administrative costs, or
- result in less DR being available for use in the RERT as the use of any baseline will be applicable to a smaller number of different types of loads as a higher frequency of achieving good or acceptable levels of accuracy is required.

4.3 Highly intermittent loads

There were very few highly intermittent loads in the program dataset, and an assessment of the relative performance of the '10 of 10' and alternative baseline methodologies was not undertaken for them.

Assessment of the performance of these loads in a RERT event is not amenable to an average demand methodology (which includes the '10 of 10' and all the alternatives assessed). To illustrate this, consider the example of a press with a maximum load of 1 MW:

- If it operates at maximum load for five minutes in a one-hour period, it will consume 83 kWh
- If it operates at a constant load of 83 kW for the entire hour it would also consume 83 kWh.

Under any of the baseline methodologies assessed in this study, these two outcomes would be seen to be the same despite the difference in their implications for the power system. In the former case, the instantaneous demand of this load that would need to be met by central generation and the grid would be 1 MW, not 83 kW. This could be of significance in a RERT situation when generation capacity is severely limited as compared to expected aggregate demand.

A baseline based on instantaneous demand (or 5-minute average demand)[26] could provide a more accurate assessment of the level of demand response delivered, however, this was not tested within this study.

4.4 Commercial loads

A total of 228 NMIs with commercial loads were analysed, 179 in Victoria and 49 in NSW. The results in the two states again differed considerably.

- The '10 of 10' method did not perform particularly well with the Victorian commercial NMIs.
  - It produced a 'good' level of accuracy in only 48 per cent of the simulated event days, and an 'acceptable' level of accuracy in 76 per cent of them.
  - The inclusion of a weather factor and consideration of the day of the week did not result in more accurate baselines.
  - This is likely a product of the fact that the Victorian commercial sample had a very skewed distribution of customer sizes. Two of the NMIs had a peak interval demand of 450 kWh, while in the other 177 the average peak interval demand was 0.45 kWh. The small sites are likely to be relatively more weather sensitive. By contrast, the two larger sites dominate the aggregate load (92 per cent) of this segment, but also exhibited significant variation in load on a half-hourly basis during event-period hours. This likely had the same effect on the accuracy of the baselines as did the highly variable industrial loads in NSW discussed above.
- The '10 of 10' method performed much better with the NSW commercial NMIs.
  - It produced a good level of accuracy in 85 per cent of the cases, and an acceptable level of accuracy in 100 per cent of the simulated event days.
  - Consideration of the day of the week and temperature with a 40 per cent adjustment factor cap[27] improved results, increasing the percentage of simulated event days with good accuracy to 100 per cent.

26 The NEM will transition to 5-minute settlement from 2021.

27 As discussed later, the increase in the adjustment factor may be as, or more, important than the inclusion of temperature and day-of-week.

Figures 4 and 5 illustrate the differences in the accuracy and bias of the various baseline approaches as applied to the commercial NMIs in Victoria and NSW.

Figure 4: Scatterplot results of accuracy and bias for VIC commercial NMIs

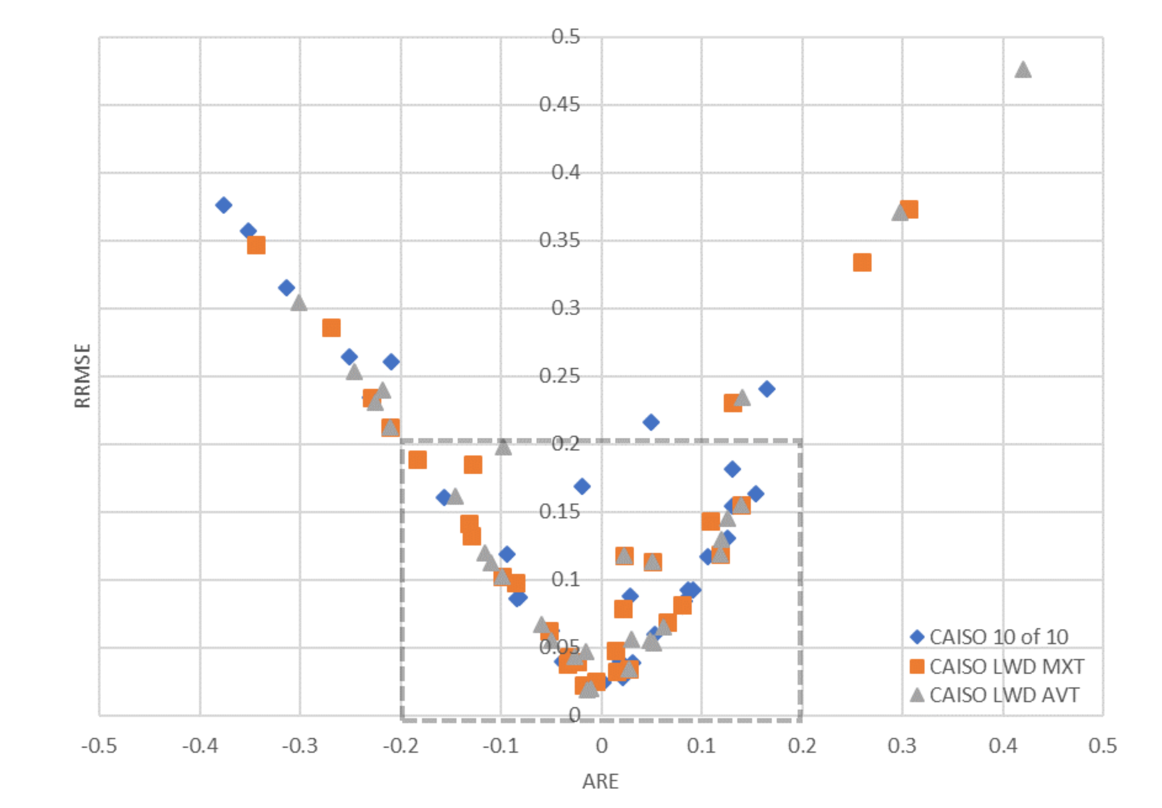


Figure 5: Scatterplot results of accuracy and bias for NSW commercial NMIs

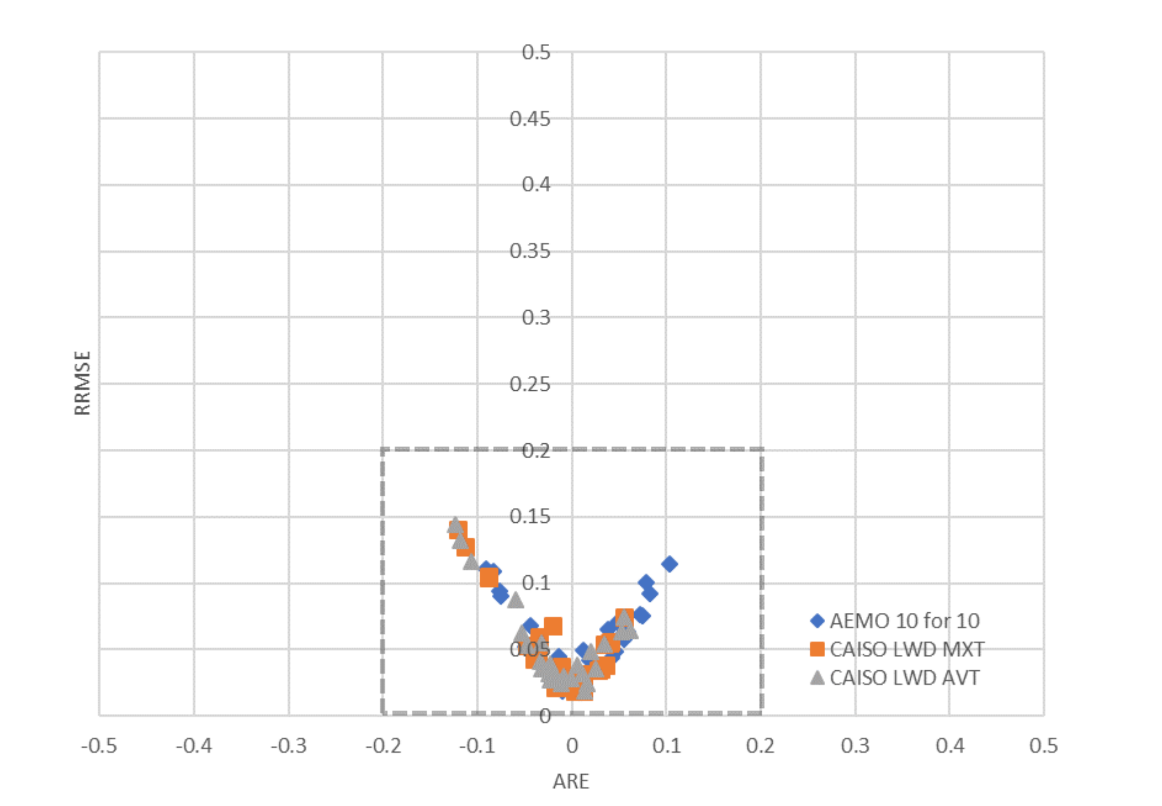


Table 4 summarises the results shown in Figures 4 and 5 in terms of the accuracy of the various baseline methodologies tested for application in the commercial sector. None of the baseline approaches produced an acceptable level of accuracy in more than 79 per cent of the simulated event days for commercial loads in Victoria. By contrast, all of the baseline approaches did so for 100 per cent of the events in NSW. However, in both states the use of a like day of the week and average temperature provided the best results.

Table 4: Percentage of simulated event days for commercial loads with ‘good’ and ‘acceptable’ accuracy using different baseline methods

Jurisdiction & Baseline method	Good accuracy (RRMSE < 10%)	Acceptable accuracy (RRMSE < 20%)
VIC commercial NMIs		
‘10 of 10’	48%	76%
Maximum temperature	45%	79%
Average temperature	40%	76%
Day of week & maximum temperature	35%	66%
Day of week & average temperature	48%	76%
NSW commercial NMIs		
‘10 of 10’	85%	100%
Maximum temperature	85%	100%
Average temperature	89%	100%
Day of week & maximum temperature	100%	100%
Day of week & average temperature	100%	100%

4.5 Loads that vary from day to day, but in a consistent pattern

Only a small number of NMIs – all of which were commercial facilities – exhibited a consistently variable load type. Retail establishments with extended trading hours on Thursday evening were the most common example of commercial facilities with a consistently variable load.

Figure 6 illustrates the differences in the accuracy and bias of the various baseline approaches as applied to the commercial NMIs in one NSW proponent's portfolio, which included several consistently variable loads.

Figure 6: Scatterplot results of the accuracy and bias for the commercial NMIs in NSW portfolio characterised by 'consistently variable' loads

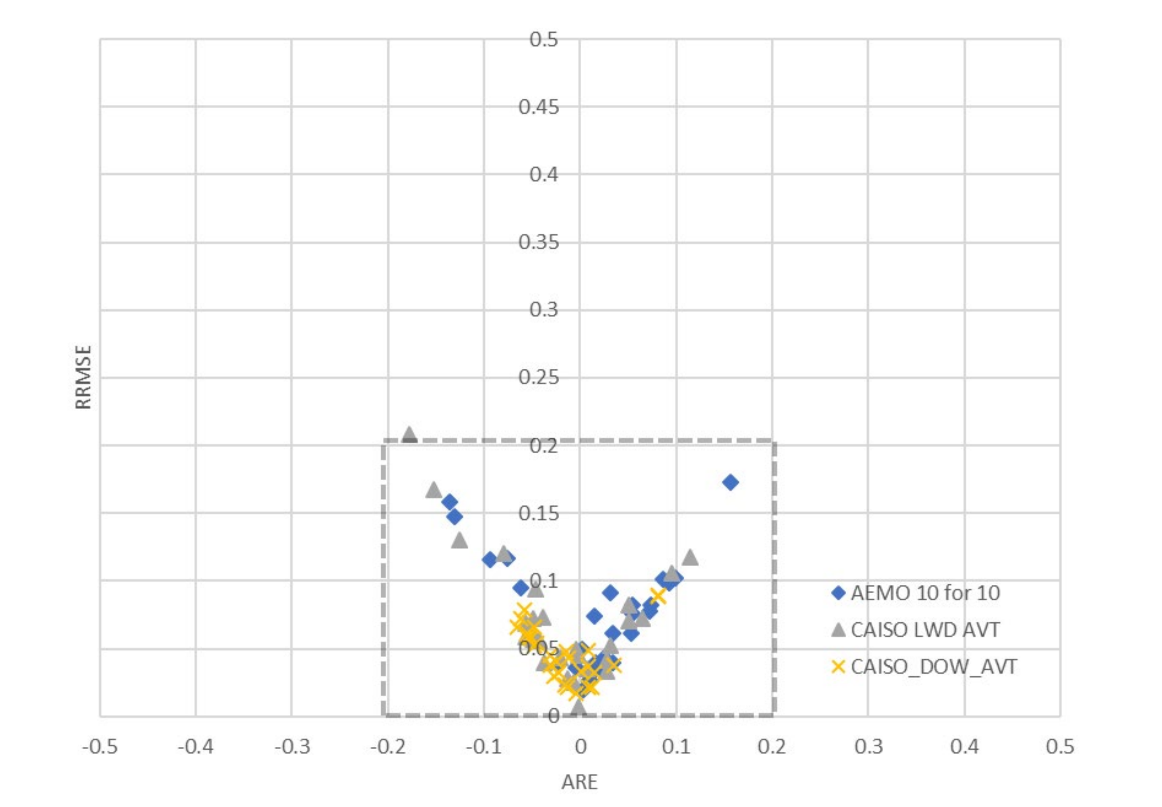


Table 5 summarises the results shown in Figure 6 in terms of the accuracy of the various baseline methodologies tested for this portfolio. It shows that the use of a 'day of week' approach and either maximum or average temperature was able to provide a 'good' level of accuracy in all simulated events.

Table 5: Percentage of simulated event days with 'good' and 'acceptable' accuracy for a NSW portfolio with consistently variable loads using different baseline methods

Jurisdiction & Baseline method	Good accuracy (RRMSE < 10%)	Acceptable accuracy (RRMSE < 20%)
NSW commercial NMIs		
'10 of 10'	76%	100%
Maximum temperature	83%	100%
Average temperature	80%	99%
Day of week & maximum temperature	100%	100%
Day of week & average temperature	100%	100%

4.6 Residential loads

Metering data on residential NMIs was only available from Victorian DR portfolios.

4.6.1 Weather-sensitive loads

Weather sensitivity is most pronounced in the residential and small commercial sectors. Figure 7 illustrates the differences in the accuracy and bias of the various baseline approaches as applied to weather sensitive residential loads in Victoria.

Figure 7: Scatterplot results of accuracy and bias for VIC residential weather-sensitive loads

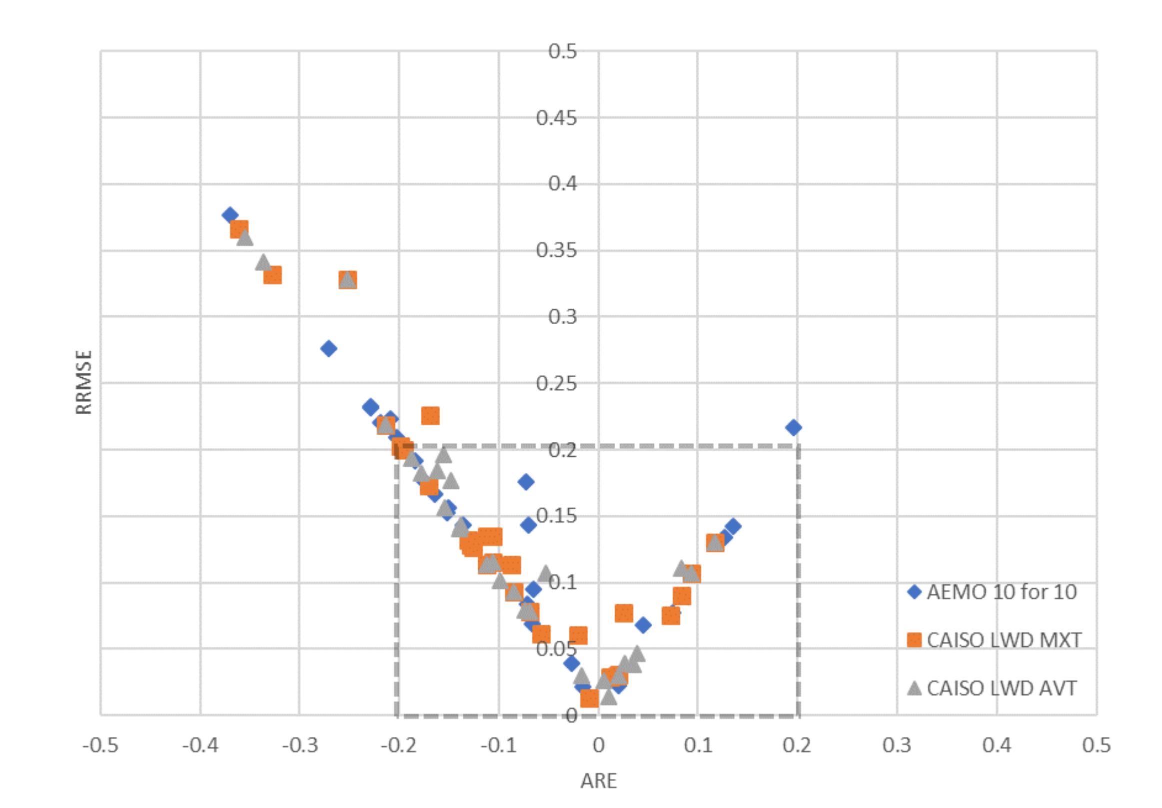


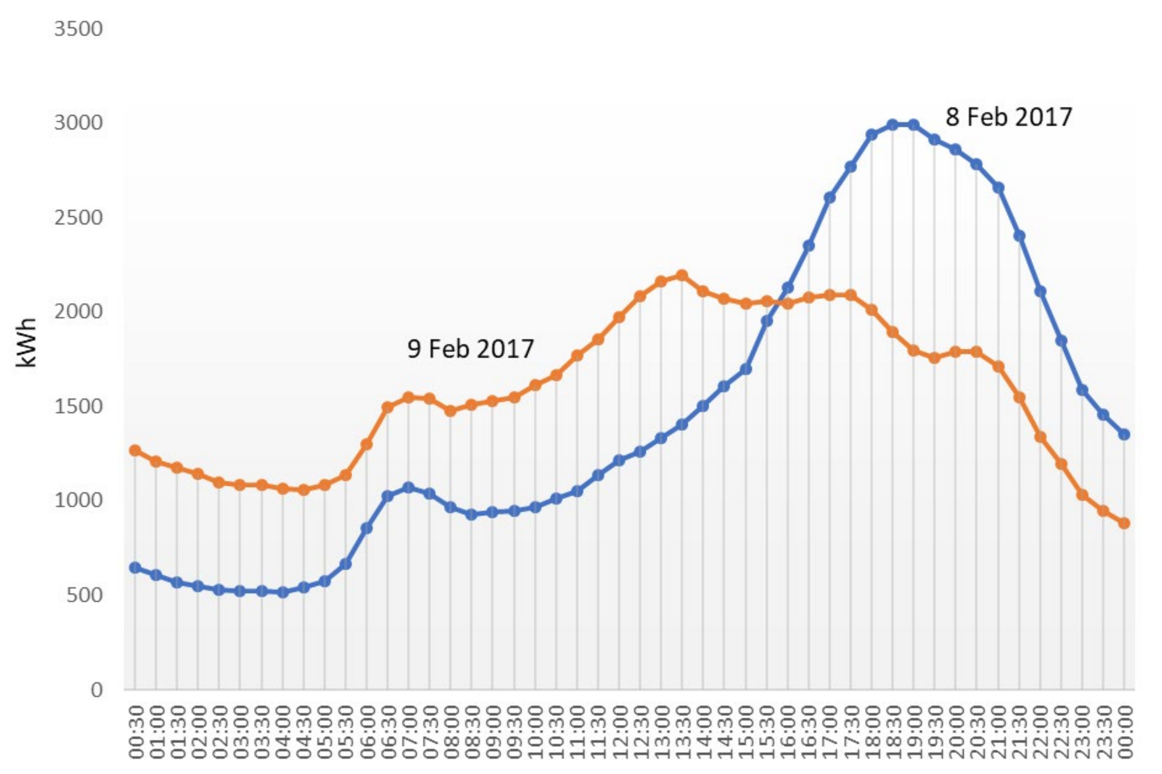
Table 6 summarises the results shown in Figure 7 in terms of the accuracy of the various baseline methodologies tested for weather-sensitive residential NMIs in Victoria. It shows that none of the methodologies tested resulted in baselines with 'good' accuracy in more than 40 per cent of the simulated events. By contrast, all of the methodologies produced an acceptable level of accuracy in at least 70 per cent of the simulated events, with the use of average temperature seeming to be a bit more effective than maximum temperature. Inclusion of consideration of the day of the week did not show material improvement in results.

Table 6: Percentage of simulated event days with 'good' and 'acceptable' accuracy for weather-sensitive residential loads in VIC using different baseline methods

Customer segment & baseline method	Good accuracy (RRMSE < 10%)	Acceptable accuracy (RRMSE < 20%)
VIC residential non-PV NMIs		
'10 of 10'	27%	70%
Maximum temperature	37%	76%
Average temperature	33%	83%
Day of week & maximum temperature	40%	77%
Day of week & average temperature	23%	80%

Figure 8 shows the aggregate load profile of non-PV residential RERT Trial participants in Victoria for the 8th (in blue) and 9th (in red) of February 2017. These profiles show that load shape (and slope) can differ substantially from day to day.

Figure 8: Load profiles of non-PV residential NMIs in VIC for 8 & 9 February 2017



The two days shown in Figure 8 were the peak demand days in the 2016-17 summer in Victoria. Both these days had similar maximum temperatures of 35°C to 36°C and average temperatures of 26.7°C to 26.8°C. The '10 of 10' approach and even a like-weather day methodology is based on the assumption that these two days would be expected to have the same (or very similar) load profiles. However, as can be seen, they do not. In particular, the load profile on the second day has much lower peak demand than the first day despite the fact that they were very similar in average and maximum temperature.[28]

These load profiles also run counter to the general view that peak demand increases over the course of successive hot days. That observation may hold true for total load in an area, even if the pattern shown here suggests that residential peak demand may be lower and occur earlier on the second (and possibly subsequent) hot days. However, the pattern has implications for the accuracy of the '10 of 10' or even like-weather day baseline approaches when applied to residential DR portfolios in periods of consecutive hot days.

4.6.2 Loads influenced by rooftop PV generation

Rooftop PV systems were only identified to be in use in the residential sector. In addition, all of the sites for which metering data was available were in Victoria where PV systems are net metered, making it impossible to disaggregate electricity consumed from the PV system from total household electricity consumption. This in turn makes it difficult to assess how the household's consumption of electricity from the grid on a RERT day differs from its non-RERT day (i.e. baseline) consumption.

28 Energy consumption on 8 February 2017 (the highest peak day) was 70,934 kWh; on the 9th it was 75,967 kWh. One possible explanation of these differences may be that air-conditioning is turned on in the afternoon of the first hot day causing a massive peak load, but then is likely to remain on overnight and throughout the next day due to the sustained high temperatures. This results in a higher level of consumption in the morning of the second day but avoids the start-up load required on the first day, thereby resulting in a much softer peak on the second day. It should be noted that the thermal resistance of the home will also play a role in determining how quickly the internal space heats up and the resulting timing of the need for air-conditioning.

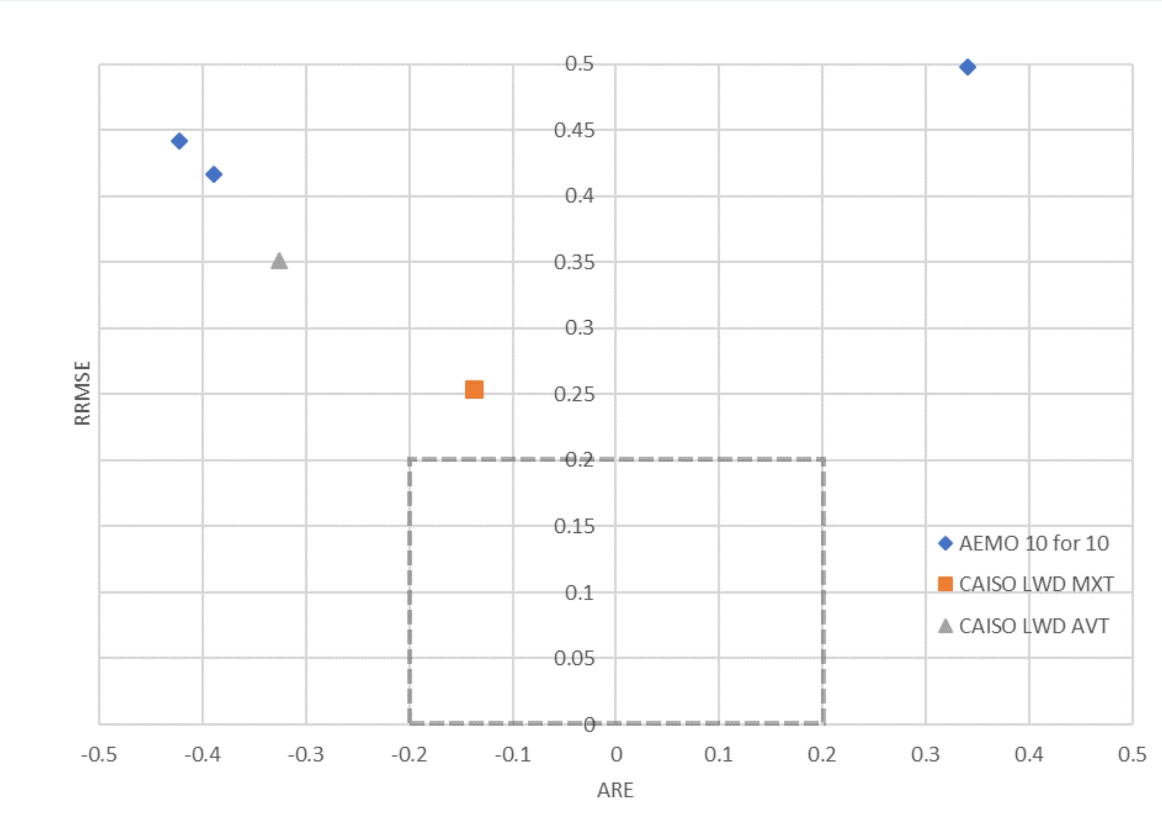
An indirect method comprised of the following steps was undertaken to estimate and thereby normalise average household consumption for PV output:

- An estimate was made of the aggregate PV generation in two two-digit postcodes.[29] The key inputs to the estimate were the PV production profile of a PV system within the two postcodes, and data from the Australian PV Institute (APVI) on the half-hourly electricity generation of the average-sized PV system installed within those postcodes.
- The resulting aggregate half-hourly PV generation was added to the metered load of the sites with PV in those two postcodes.
- This was assumed to produce the gross consumption of these residences (as any export would have been accounted for as negative consumption in the net metered PV load).
- The '10 of 10' and alternative baseline approaches were then applied to this gross assumption.

None of the baseline approaches that were tested were found to produce good or acceptably accurate baselines for the residential PV segment in any of the simulated events. As can be seen by the very few data points in Figure 9 below, very few of the simulated events produced RRMSE values lower than 0.5.

As noted above, this result is very much a product of the inability to account for total PV output where net metering is used. It is likely that gross metering of the PV contribution (which would allow accurate assessment of the total consumption of electricity within the house) would produce baseline accuracy results closer to those discussed above for houses without rooftop PV systems.

Figure 9: Scatterplot results of accuracy and bias for VIC non-PV residential NMIs



29 30XX and 31XX, which cover Melbourne and the greater Melbourne suburban area.

Table 7 summarises the results shown in Figure 9 in terms of the accuracy provided by the various baseline methodologies tested for residential properties with rooftop PV systems in Victoria.

Table 7: Percentage of simulated event with 'good' and 'acceptable' accuracy for VIC residential properties with rooftop PV systems using different baseline methods

Customer segment & baseline method	Good accuracy (RRMSE < 10%)	Acceptable accuracy (RRMSE < 20%)
VIC residential PV NMIs		
'10 of 10'	0%	0%
Maximum temperature	0%	0%
Average temperature	0%	0%
Day of week & maximum temperature	0%	0%
Day of week & average temperature	0%	0%

4.7 Impact of a larger adjustment factor

All of the alternative baselines included an increase in the adjustment factor cap from the +/-20 per cent used in the '10 of 10' approach to +/-40 per cent. The adjustment factor cap places a limit on the amount by which the baseline can be adjusted to account for any difference in the consumption of a facility (or a portfolio) in the lead-up to dispatch of DR on an event day when compared to that same time period in the qualifying days that are used to create the baseline. Where the difference in consumption in that time period between event days and qualifying days is large, a smaller cap will result in a lower baseline and an under-calculation of the amount of demand response delivered.

The table below shows the percentage of total simulated event day observations in which the 20 per cent and 40 per cent adjustment factor caps allowed the full difference between consumption in the pre-event period on the event and qualifying days to be reflected in the adjusted baseline. The 20 per cent cap caused a moderate level of restriction in the adjustment of baselines for the industrial and commercial sectors, though slightly more in the commercial sector, and a significantly larger proportion of the baselines for residential facilities. The use of the 40 per cent cap resulted in a material reduction in the proportion of baselines whose adjustments were restricted in the commercial and some of the industrial portfolios, but a lesser reduction in the residential portfolios.

The outlier in Table 8 is the NSW industrial sector, in which about the same number of the baselines were restricted with either the 20 per cent or the 40 per cent adjustment factor cap. This is probably a by-product of the high level of half-hourly load variability of that group (as discussed in Section 4.2).

Table 8: Proportion of simulated event days in which the cap restricted the adjustment of the baseline

Jurisdiction / sector	10 of 10 (+/- 20% cap)	Alternative baselines (+/- 40% cap)
Industrial loads		
Victoria	14%	None
New South Wales	33%	37% to 47%
Commercial loads		
Victoria	30%	0% to 3%
New South Wales	43%	None
Residential loads		
Victoria	77%	57% to 77%

Tests were not run on the ‘10 of 10’ approach with a 40 per cent (rather than the 20 per cent) adjustment factor cap. As a result, the impact of the larger cap on the accuracy of the baseline has not been quantified. However, given that actual consumption on the simulated event days was known, and the accuracy test was based on how well the baseline could predict that consumption, it is reasonable to expect that restriction of the baseline would reduce accuracy.

On the other hand, the use of a larger adjustment factor cap creates an opportunity for gaming. If a DR provider knows that a RERT event will be called, they could theoretically increase their consumption in the pre-event period which would have the effect of increasing their adjusted baseline and therefore increasing the DR calculated from it. The larger the adjustment cap, the greater the opportunity for gaming of this sort. This can be reduced by other program measures, for example the amount of notice given for an event, and the length and proximity of the adjustment window to the event period. There is also the question of whether the pay-off from such gaming would be worth it given the monetary cost of the artificial increase in consumption and the risk that the event might not be called in the end.

5. ALTERNATIVE APPROACHES TO BASELINING

5.1 Approaches identified by proponents in the ARENA program

Three of the proponents developed alternative approaches for calculating customer baselines. These have not been assessed as part of this study but are discussed below.

5.1.1 AGL

AGL developed and used a methodology based on the concept of anchoring, which averages the usage at a particular time of the day for days of a similar temperature over the last five weeks (differentiating usage on weekdays and weekends) and anchors it to the actual consumption before and after the event. It is calculated using the following steps:

- Generation of a site level forecast based on regression of the previous five weeks net load (load minus solar) excluding any controlled load channels against temperature, time of day and workday/non-workday. The regression of the previous five weeks net load is used to identify the best "shape" that corresponds to the event day.
- De-biasing by comparing the previous seven days' forecasts against the actuals for the same time of day in the event period and adjusting the event period baseline forecast.
- Anchoring the predicted consumption outside the event period to the actual consumption on that day, based on smoothed consumption either side of the event period.

AGL stated that its approach allows the adjustment period to better account for the specific temperature on the day and that it is suitable for both PV and non-PV customers, thereby improving the accuracy of this approach as compared to the '10 of 10'.

5.1.2 United Energy

United Energy proposed an alternative baseline approach that retains the '10 of 10' method with an adjustment period, plus the additional parameters:

- Extends the period of qualifying days to two years prior to the event day.
- Selects the 10 reference days based on like maximum temperatures (i.e. within 4°C of the event day maximum temperature) at a defined Bureau of Meteorology site.
- Uses the hour immediately preceding the event as the adjustment period (shorter than the 3-hour period used in the '10 of 10' method).
- Retains the 20 per cent adjustment cap.

5.1.3 Zen Ecosystems

Zen Ecosystems used a linear baseline as an alternative to the '10 of 10' approach in assessing the performance of its behavioural residential DR portfolio on a test day characterised by a very high temperature.

The linear baseline is constructed by drawing a straight line from the portfolio's average half-hourly metered consumption approximately 1 hour prior to the start of the DR event until approximately 1 hour after the close of the event. Note that this does not use previous days – the baseline is constructed using only the actual metered data of the portfolio on the event day. This approach has some of the characteristics of anchoring, in that it effectively averages variations between the hour before and the hour after the event.

5.2 Approaches identified in other jurisdictions

PJM and South Korea use alternative baselining approaches in their demand response programs. Their approaches use the same basic components as those used in the CAISO '10 of 10' methodology, though the specifics of the components differ from those in the '10 of 10' approach. These approaches described in this section are included to provide information on other options.

5.2.1 PJM

PJM uses a different baselining approach for DR that is used in the energy market or to provide ancillary services, as compared to the approach it uses for DR deployed to provide emergency capacity.

PJM's baseline method for DR deployed in the energy market or to provide ancillary services has the following features:

- **Selection Window** – The selection window is 45 calendar days.
- **Qualifying Day** – The three, five or seven (depending on the baseline method selected) most recent non-event days matched by either the day of the week or the weather according to defined criteria.
- **Reference Days** – Either the three or five middle demand days. The average of the load profiles on the reference days is the proxy baseline, called the 'unadjusted baseline'.
- **Adjustment Period** – Three hours with a one-hour buffer prior to the event.
- **Adjustment Limit** – 20 per cent is the maximum amount that the baseline can be varied as a result of differences during the adjustment period between the actual load on the day of the event and the baseline load.

PJM has also developed a separate approach for DR that is used to provide emergency capacity, which could provide some useful transferability to the RERT. For participation in the PJM Capacity Market, participants must have sufficient metering and be capable of 1 or 5 minute settlement. An anchoring type of approach with real time metering is used to measure the DR contribution. Participants must be able to demonstrate an actual change in demand for the defined period of the capacity call.

5.2.2 South Korea

In South Korea, demand response participants can choose between two baseline approaches:

- **The '6 of 10' method** – The average of the middle six days (by load) of the most recent 10 qualifying days (from up to the previous 20 days) before the event day. The two highest load and the two lowest load days from the 10 days are discarded.
- **The '4 of 5' method** – The average of the four highest load days selected from the previous five qualifying days (from up to the previous 10 days).

For each method there are two optional adjustments allowed:

- **Unusual days** – Proponents are allowed to remove any day ('unusual day') where the average load for that day is less than 75 per cent or over 125 per cent of the highest average load day for the reference period (note that for residential customers only the lower limit is used). If the number of qualifying days falls below the number required, the most recent 'unusual day' is used in the calculation.
- **Symmetric Additive Adjustment** – This is a similar approach to that used by AEMO where the baseline can be adjusted based on a comparison of the load for a three-hour period (from four hours to one hour) before the commencement of the demand response. This is applied as an addition or subtraction from the baseline.

This approach allows the baseline method to be tailored to the site to provide the most accurate and beneficial results for the site and therefore for the market.

6. CONCLUSIONS

The results of this analysis suggests that the '10 of 10' baseline methodology currently used in the RERT is adequate for certain types of loads, particularly those of larger commercial and industrial customers whose energy consumption is relatively similar from day to day and not particularly weather sensitive.

However, the '10 of 10' methodology does not – and was not designed to – function particularly well in predicting the consumption of loads that do not exhibit a high level of consistency, particularly in the shape but also in the level of their electricity consumption from day to day. This is especially problematic in the case of the RERT which is often (though not always) activated on days characterised by very high levels of demand, which also changes the demand profiles of many types of customers.

Examples of these types of loads that were present in the portfolios of the proponents in the RERT Trial analysed in this study included:

- Residential customers with and without rooftop PV systems. In the case of residences without PV systems, this is likely the impact of weather sensitivity and household occupancy patterns that vary from day to day. In the case of residences with rooftop PV systems, this is likely to do with variations in the amount of energy available from the PV system from day to day as well as the factors affecting residences without PV.
- Some commercial loads, particularly those with a high degree of weather sensitivity or operational factors that reduce the similarity of the consumption and consumption profile from day to day.
- Larger loads that exhibit a high degree of variability in their electricity consumption that is driven by internal operational rather than external factors.

Although no examples were included in this analysis, it is likely that the '10 of 10' baseline methodology would also not function well in the case of commercial loads with PV systems whose installed capacity is a material proportion of the average load of the facility itself.

Several modifications to the '10 of 10' methodology were tested for their ability to provide better predictions of the consumption for portfolios comprised of different types of customers and loads. These modifications include how the qualifying days used to construct the baseline are selected, such as selecting qualifying days based on their similarity to the event day in terms of their:

- maximum temperature
- average temperature
- maximum temperature and day of the week
- average temperature and day of the week.

Each of these alternative methods include an increase in the cap on the amount by which the baseline can be adjusted based on the difference in consumption levels prior to the commencement of RERT activation on the event day as compared to that same time in the baseline. The increase in the cap was a 40 per cent difference, as compared to the 20 per cent cap used in the '10 of 10' approach.

Findings regarding the performance of these modified versions of the '10 of 10' approach were as follows:

- None of the modified versions provided anything more than exceedingly marginal improvements in as compared to the '10 of 10' approach for portfolios comprised of large industrial customers.
- Not surprisingly, the use of day of the week in combination with either maximum of average temperature was shown to improve results in the case of some commercial loads – primarily with those whose operating schedules show consistent variation by day of the week.
- Although several of these modifications resulted in reasonable improvement in the ability to predict the consumption of residential customers without PV systems on a RERT-type day, even with those improvements the frequency that any of the baselines tested provided good or acceptable accurate predictions of that consumption was still too low to be considered useful.

- None of the modified approaches improved on the ability of the ‘10 of 10’ baseline to predict the consumption when applied to residential customers with PV systems. None of the methods functioned at a level that would warrant any further consideration of their use.

Other approaches may offer better alternatives for the types of loads for which the ‘10 of 10’ baseline does not adequately predict event-day consumption. These include anchoring or the use of control groups. Anchoring assesses the shape of consumption of the facility on days of like temperature in the past and uses that shape and the pre- and post-period consumption of the facility on the event day to construct the baseline.

A control group is a group of customers whose consumption on event days can be assumed (or has been shown) to be similar to that of the customers providing DR. The consumption on the day of the DR event of the control group is then assumed to represent what the consumption of the DR customers would have been, and the difference between the consumption of the two groups is taken to represent the amount of DR delivered.

The main issues in the construction of a control group are considerations regarding how similar the customers in the control group are to the DR customers, and the size of the control group needed to provide a suitable level of statistical validity in the comparison between the consumption of the control and DR groups.

For any comments or questions on this report, please contact knowledge@arena.gov.au