THE UNIVERSITY
of ADELAIDE

Review of FORECAST ACCURACY METRICS

For Australian Energy Market Operator - 2023

GROUP OF EIGHT AUSTRALIA

make history.

Table of Contents

Citation: Bean, N.G., Ross, J.V. (2023) Review of forecast accuracy metrics for the Australian Energy Market Operator. The University of Adelaide, Australia.

Executive Summary

The Australian Energy Market Operator (AEMO) annually produces forecasts for each region of operational energy consumption, extreme half-hourly demands, and available supply. AEMO reports on the forecast accuracy in the annual Forecast Accuracy Reports (Annual Reports). The Annual Reports build confidence in AEMO forecasts and *inform the continuous improvement of the forecasts.*

The University of Adelaide was engaged to review AEMO's current forecast metrics and forecast accuracy reporting methodology.

This review follows on from the 2019 Review of Forecast Accuracy Metrics, also by The University of Adelaide, and assesses the forecast accuracy metrics that are described in the "Forecast Accuracy Report methodology" paper dated August 2020 and presented in the 2020, 2021 and 2022 Annual Reports.

We note that many of the recommendations of the 2019 Review have been implemented and have significantly improved the communication of the forecast accuracy.

We found the "Forecast Accuracy Report methodology", dated August 2020, to be an excellent description of the relevant issues and approaches and the quality of the accuracy reporting methodology and the Annual Reports to be very high. We commend AEMO on their standards and progress. A very brief summary would be to "continue doing what you are doing".

Nonetheless, we provide 26 recommendations from our review that we believe will enable AEMO to further improve their current practice. These recommendations are based on the high-level information available to the review team. AEMO will need to assess the feasibility and benefits of implementing each recommendation. Specific details around what information should be presented or prioritised is to be determined internally by AEMO and informed by stakeholders.

We have structured our report (largely) in line with the structure of the Forecast Accuracy Report methodology", dated August 2020 and the Annual Reports. The recommendations can also be categorised as:

	Recommendations
Commendations – continue with current practice	1, 2, 3, 9, 13, 16
Recommendations that concern presentation only	5, 10, 11, 12, 18
Recommendations that will require some development	4, 6, 14, 15, 17, 19, 20, 21, 22, 23, 24, 25, 26
Recommendations that are more strategic in nature, which we recognise may not be achievable in the short-term, but we believe provide useful long-term goals	7, 8

List of Recommendations

2 General Recommendations

Recommendation 1: Continue with the use of the forecast categories and reporting methodologies described in Table 4 of the “Forecast Accuracy Report methodology”, dated August 2020.

Recommendation 2: Continue with the use of forecast accuracy reporting as a tool to drive improvements in the forecasting methodology.

Recommendation 3: Continue with the use of Definition 3 for percentage error. It is the more easily interpreted definition given the framing of the report as assessing the accuracy of the forecast against the actual. Ensure that this framing is used consistently throughout.

Recommendation 4: Consider providing information for each row in Table 1 of each Annual Report indicating the assessability of that metric, using the three categories defined in Section 2.2.1 of each Annual Report.

Recommendation 5: Rewrite the description of a box plot to prevent potential confusion between outliers and the maximum/minimum.

Recommendation 6: Consider the introduction of enhanced representations of weather in the descriptions and the models to enable a more rigorous analysis of accuracy.

Recommendation 7: Consider opportunities to benchmark the accuracy of the forecasts against other organisations.

Recommendation 8: Consider introducing 2-year and 4-year assessments of the accuracy of certain key elements of the forecasts in the annual reporting process.

3 Operational energy consumption forecasts

Recommendation 9: Continue with the use of percentage error and percentage impact on forecast of total consumption and with the use of tables and waterfall diagrams to represent them.

Recommendation 10: Replace equation 2 on page 17 of the “Forecast Accuracy Reporting methodology” paper dated August 2020 with the equation:
$error\ from\ forecast\ component = input\ coefficient.\,(input\ \boldsymbol{forecast} - input\ \boldsymbol{actual})$.

Recommendation 11: Reorder all the waterfall diagrams to be consistent with the new equation 2 so that each waterfall figure starts with the Actual and presents all the component errors in the forecast that are required to reach the Forecast.

Recommendation 12: Ensure that the order and labels of the components in all waterfall diagrams and associated tables are consistent. Consider if it is appropriate to present the three supply-side components (that require a reverse of sign) first in each table and waterfall diagram.

4 Extreme demand forecasts

Recommendation 13: Continue the use of a discussion-based approach and the use of figures that provide meaningful information about the distribution and drivers of the forecast.

Recommendation 14: Retain Figure 15, (and its associated versions for each region) in future Annual Reports. Review the choice of driving parameters that are displayed and how each parameter is presented.

Recommendation 15: Consider whether it would be more appropriate to provide the monthly maximum demand figures based on only 10%, 50% or 90% POE traces, or provide them based on the combination of the 10%, 50% and 90% POE traces all together.

5 Supply forecasts

Recommendation 16: Continue using the figures for total availability and component generation for each region.

Recommendation 17: Restricting the graph to the central 95% is a commonly used and entirely appropriate approach. However, consideration could also be given to other approaches that are designed to achieve a similar degree of interpretability.

Recommendation 18: Forecast and actual generation count and capacity tables should be restructured so that forecasts are provided to the left of the actuals and that the comparison column is calculated as (forecast – actual) so that the final column follows the generic definition of percentage error.

Recommendation 19: Consider providing an equivalent analysis of the accuracy of supply forecasts in the most important supporting regions based on the top 10 hottest days in the supported region.

Recommendation 20: Consider whether the accuracy of VRE generation forecasts could be presented consistently on a generation basis through both the forecast and reported data.

Recommendation 21: Consider modifying the trigger categories into disjoint categories (eg ">=$300/MWh AND <$500/MWh" for the lowest trigger category). Further, consider reducing the number of categories to ensure sufficient events in each (disjoint) category while maintaining signal and interpretability.

Recommendation 22: Consider reporting the accuracy of the demand side participation forecasts by comparing the forecast distribution with the observed distribution, this could be achieved using side-by-side box plots or violin plots, for example.

Recommendation 23: Investigate ways to clarify the presentation in the "DSP response during reliability events" section to assist the reader's understanding.

6 Potential Modelling Improvements

Recommendation 24: Consider the introduction of further industry segmentation to improve consumption forecasting.

Recommendation 25: To increase the robustness of the forecast process, increase the number of weather years that are incorporated. This could be achieved by using more historical data or, preferably, by making use of synthetic weather years.

Recommendation 26: Consider upgrading the accuracy and level of assurance of the assumed scale factors in the Potential adjustment – voluntary load reductions feature.

1. Introduction

The Australian Energy Market Operator (AEMO) annually produces forecasts for each region of operational energy consumption, extreme half-hourly demands, available supply and reliability. These forecasts are subject to high levels of scrutiny by various stakeholders.

Forecast accuracy is reported annually (as per clause 3.13.3(u) of the National Electricity Rules) in the annual Forecast Accuracy Reports (Annual Reports). The purpose of the Annual Reports is to build confidence in AEMO forecasts and help inform the continuous improvement of these forecasts. The Annual Reports have a wide audience including government, decision-makers, and the general public; as such they must present metrics that are appropriate and accessible.

This report assesses the forecast accuracy metrics that are described in the “Forecast Accuracy Report methodology” paper dated August 2020 and presented in the Annual Reports. For the purpose of the Annual Reports, we have considered in detail the 2020, 2021 and 2022 versions since those are the Annual Reports that have been prepared following the 2019 Review of Forecast Accuracy Metrics by The University of Adelaide.

It is worth noting that the reliability forecasts are included in the Annual Reports for information only and are not presented for the purposes of assessing forecast accuracy. Therefore, we have not considered the reliability forecasts in this report.

2. General Recommendations

The quality of the accuracy reporting methodology and the Annual Reports is very high and we commend AEMO on their standards and progress. We found the “Forecast Accuracy Report methodology”, dated August 2020, to be an excellent description of the relevant issues and approaches. We found Table 4 to be an excellent summary of the different types of forecasts and the most relevant ways of assessing accuracy in each circumstance.

Recommendation 1: Continue with the use of the forecast categories and reporting methodologies described in Table 4 of the “Forecast Accuracy Report methodology”, dated August 2020.

AEMO has a clear and living philosophy of using the forecast accuracy reporting mandate to not only keep its stakeholders informed on the accuracy of its forecasts but also to drive improvements to its forecasting methodology. This is a commendable approach and it is rewarding to see the direct benefits of this philosophy in changes to forecast methodology.

Recommendation 2: Continue with the use of forecast accuracy reporting as a tool to drive improvements in the forecasting methodology.

Percentage Error on page 16:

Definitions 1 and 3 are both equally statistically accurate and so the relevant choice should be made according to ease of interpretation. The reporting methodology is framed around assessing the accuracy of the forecast against actual (or estimated actual) quantities. Under this framing, Definition 1 is counter-intuitive since a *positive* error would indicate a forecast that is *lower* than the actual. We note that we used Definition 1 in our previous report, because that was the definition that was being used by AEMO at the time. We endorse the change to using Definition 3, since it means that a *positive* error now indicates a forecast that is *higher* than the actual.

It is then important that all discussions are framed as assessing the accuracy of the forecast against actual (or estimated actual) quantities with all language and calculations consistently following this framing. This consistent framing should always reference/present the forecast first, followed by the actual and any description should be written in terms of how the forecast compares to the actual. We noted a number of situations where this has not been followed, many of which we will note in the appropriate sections.

Recommendation 3: Continue with the use of Definition 3 for percentage error. It is the more easily interpreted definition given the framing of the report as assessing the accuracy of the forecast against the actual. Ensure that this framing is used consistently throughout.

Section 3.3 of each Annual Report:

We support the notion of averaging the two independent forecasts, given that they are based on the same set of underlying assumptions.

Table 1 of each Annual Report:

We found the traffic light summary in Table 1 of each Annual Report to be quite informative. We thought it might be made more informative by introducing some explicit indicator (perhaps shading) for each row indicating the assessability of that metric. Presumably, it would make sense to use the three categories defined in Section 2.2.1 of each Annual Report.

Recommendation 4: Consider providing information for each row in Table 1 of each Annual Report indicating the assessability of that metric, using the three categories defined in Section 2.2.1 of each Annual Report.

Definition of "box plot" in each Annual Report:

The description of a box plot does not follow standard statistical language. If the boxplot allows for "outliers" to be displayed, then the end of the "whiskers" are not necessarily the maximum and minimum but are defined in a more sophisticated way. As such, we don't recommend the use of the terms Maximum and Minimum, for if that were the case there could be no outliers. There are three possible approaches that we believe would be preferable to the current approach:

1. Present the whiskers as the true maximum/minimum, label them as such and have no outliers.
2. Use the current whiskers that allow for outliers and label them as the lower/upper whiskers, leaving their technical definition unstated, but referenced.
3. Use the current whiskers that allow for outliers and label them as the lower/upper whiskers, while also providing their technical definition.

Recommendation 5: Rewrite the description of a box plot to prevent potential confusion between outliers and the maximum/minimum.

Enhanced representation of weather in the descriptions and the models

The descriptions of the weather conditions in the explanation of the extreme demand forecasts were very interesting. The current descriptions are largely based on air temperature at a particular time and location. However, the descriptions were often elaborated upon to provide a more nuanced picture of the weather on that day across that region. This raises the question as to whether there may be more relevant weather measures that could be used to enhance both the explanation and also the underlying forecast process for extreme demand and other weather-impacted forecasts. This would lead to a more rigorous analysis of accuracy.

Specific ideas include:

- Use of a demand-weighted spatial average of temperature in each region. This is probably most likely to be relevant in Queensland where notable demand stretches across a very large region with quite different conditions.
- Use of the concept of a heating/cooling load up to that point in time on that day as often the accumulation appears to be more important than the actual temperature at that time.
- Use of the wet-bulb or apparent temperature to include the effects of humidity.

Recommendation 6: Consider the introduction of enhanced representations of weather in the descriptions and the models to enable a more rigorous analysis of accuracy.

Benchmarking

Stakeholders would have increased confidence in the AEMO forecasts if their accuracy could be benchmarked against the forecast accuracy of similar organisations. We are not aware of any other jurisdictions with the same requirements in terms of forecasting and reporting on the accuracy of those forecasts. However, it is worth considering whether there are any other related forecasts against which a benchmarking comparison could be performed. One possible example is the consumption forecasts by the Distribution Network Service Providers; this possibility has not been investigated and so there might be reasons why such a comparison would not be useful.

Recommendation 7: Consider opportunities to benchmark the accuracy of the forecasts against other organisations.

Accuracy reporting of multi-year forecasts

We understand that some of the forecasts extend up to 10 years into the future. We also note that the only accuracy metrics that are reported are based on the accuracy of the first year of each forecast (except for Figure 7 in each Annual Report relating to operational energy consumption). We understand in such a dynamic industry, hence requiring a forecasting methodology that is regularly being updated to seek improvements in accuracy and capture structural changes, that looking back at the forecasts made 10 years ago to assess their 10-year accuracy may seem to be rather meaningless.

We also understand that the 2-year and 4-year timeframes are particularly important, but for the reason above do not believe that it would make sense to repeat the entire accuracy analysis at 1, 2 and 4 years. Instead, we believe that it would be worthwhile introducing a 2-year and 4-year analysis of some of the most relevant metrics, for example, annual operational energy consumption, 10% POE extreme demand and outage rates. Lower levels of accuracy, than for the 1-year analysis, would be expected due to the continuous improvement of the forecasting methodologies and any structural/policy changes.

Recommendation 8: Consider introducing 2-year and 4-year assessments of the accuracy of certain key elements of the forecasts in the annual reporting process.

3. Operational energy consumption forecasts

The presentation of the accuracy of the operational energy consumption forecast through the presentation of the accuracy of the component forecasts is meaningful and informative. The use of percentage error, percentage impact on forecast of total consumption, tables and waterfall diagrams is appropriate.

Recommendation 9: Continue with the use of percentage error and percentage impact on forecast of total consumption and with the use of tables and waterfall diagrams to represent them.

Waterfall diagrams and associated tables and discussions

Because each waterfall diagram, and its associated table, includes demand-side errors and supply-side errors, underestimates in the forecast of a demand-side component will have to appear with the opposite sign to an underestimate in the forecast of a supply-side component. For example, consider Table 8 and Figure 7 of the "Forecast Accuracy Report 2022": Heating Degree Days and Other Non-Scheduled Generation are both underestimated, and Heating Degree days appears as a positive and Other Non-Scheduled Generation appears as a negative. This means that there is no absolute determination as to which way to present each waterfall diagram, the associated table and equation 2 on page 17 of the "Forecast Accuracy Reporting Methodology" paper dated August 2020, and it comes down to a matter of judgement about interpretability.

After much consideration, we believe that the most important factor should be consistency in the approach exemplified by the definition of percentage error, which considers Forecast – Actual, not Actual – Forecast. The fact that most components in the waterfall diagram are demand side (Cooling/Heating Degree Days, Connections Growth, Large Industrial Loads, Network Losses, and Auxiliary Load) and only three are supply-side (Rooftop PV, PV Non-Scheduled Generation, and Other Non-Scheduled Generation) is fortuitous as it means that only 3 components will have the sign of their entry in the table reversed in the figure. As these 3 components must be treated carefully, we think they should be presented first (if possible) in the table and the waterfall diagram and their change in sign should be noted.

The resulting waterfall diagram would also have to be reversed and so would start with the Actual and display all the component differences (ie errors in the forecast) that are required to result in the Forecast.

Recommendation 10: Replace equation 2 on page 17 of the "Forecast Accuracy Reporting methodology" paper dated August 2020 with the equation:
$error\ from\ forecast\ component = input\ coefficient.\ (input\ \boldsymbol{forecast} - input\ \boldsymbol{actual})$.

Recommendation 11: Reorder all the waterfall diagrams to be consistent with the new equation 2 so that each waterfall figure starts with the Actual and presents all the component errors in the forecast that are required to reach the Forecast.

Recommendation 12: Ensure that the order and labels of the components in all waterfall diagrams and associated tables are consistent. Consider if it is appropriate to present the three supply-side components (that require a reverse of sign) first in each table and waterfall diagram.

4. Extreme demand forecasts

The forecasting of extreme demand is always going to be challenging because you are dealing with the tails of a distribution which are fundamentally more variable than the more predictable measures, such as the mean. Forecasting of minimum demand is significantly harder than forecasting of maximum demand since you lose some of the benefits of the Law of Large Numbers and factors that generally might be quite minor can suddenly be quite significant.

The extreme demands are therefore highly variable and so a discussion-based approach to reporting the accuracy is highly appropriate. The figures displaying the half-hourly time series and the comparisons with the forecast distributions of both extreme demand and the driving factors are highly meaningful and informative.

Recommendation 13: Continue the use of a discussion-based approach and the use of figures that provide meaningful information about the distribution and drivers of the forecast.

Extreme demand events

Using New South Wales as an example, consider Figure 15 which displays the forecast distribution, and the actual result, across six different driving parameters. We think this figure is very instructive and support its retention in future Annual Reports.

However, we recommend that the driving parameters that are displayed should be reviewed to ensure that the maximum information is able to be gleaned. For example, the day of the week is generally not very informative since the weekday probabilities are usually all very similar, and also for the weekend probabilities. Perhaps it would be better to display these as the total probability associated with the typically understood “working” days and the total probability associated with “non-working” days. Further, the subfigure for losses is generally not very informative since losses are effectively a known proportion of the total demand. A different, potentially more informative, driving parameter could be displayed in its place – perhaps a new and informative climate parameter.

Recommendation 14: Retain Figure 15, (and its associated versions for each region) in future Annual Reports. Review the choice of driving parameters that are displayed and how each parameter is presented.

Monthly demand trace forecast performance

On page 20 of the “Forecast Accuracy Reporting Methodology” paper dated August 2020 there is an explanation of Figure 33 which presents box and whiskers plots of the monthly maximum demand from the simulations overlaid by the most recent observed actual monthly maximum. The box and whiskers are developed from the 10% and 50% POE traces generated as part of the forecasting process. Similar figures are presented in each Annual Report for each region.

If the box and whiskers plots contained only 10% POE traces, then we would have a solid intuition as to how to interpret such a figure: you would expect about 10% of actuals to lie above the 10% POE traces, and the majority to lie below. The precise comparison is hard since it is not true that you should expect 10% of actuals to lie above the extreme **maximum** 10% POE trace.

Similarly, if the box and whiskers plots contained only 50% POE traces, then we would expect about 50% of actuals to lie above the 50% POE traces, and 50% to lie below. Finally, if the box and whiskers plots contained only 90% POE traces, then we would expect about 90% of actuals to lie above the 90% POE traces, and only 10% to lie below. Again, the precise comparisons are not immediately obvious.

However, by combining the 10% and 50% POE traces together into each box and whiskers plot means that we have no meaningful interpretation of the figures. Given that the trace generation method does not produce the entire distribution but is explicitly targeted at only these 3 POE levels, perhaps, if the 10%, 50% and 90% POE traces were all combined into each box and whiskers plot, then this would be the best available representation of the entire distribution of the monthly maximum. With that understanding it would be reasonable to interpret the figure in the usual way: actuals either above the upper whisker or below the lower whisker, should only be seen rarely.

Recommendation 15: Consider whether it would be more appropriate to provide the monthly maximum demand figures based on only 10%, 50% or 90% POE traces, or provide them based on the combination of the 10%, 50% and 90% POE traces all together.

5. Supply forecasts

Forecasting supply availability is crucially important but also highly challenging given the possibility of major outages in individual, highly-significant generators and the uncertainty of the timing of operation of planned/newly-installed generators. The figures used are effective at representing that information on the most important days.

Recommendation 16: Continue using the figures for total availability and component generation for each region.

Supply availability

Restricting the graph to the central 95% is a commonly used and entirely appropriate approach, as one must remove some simulations to enable the figure to be readable. Otherwise, the figures risk showing such a broad band, due to the fact that the most extreme edges of all simulations are highly volatile, that the figure loses interpretability. Maintaining interpretability is essential, but other methods that are designed to achieve this same end could also be explored, for example, shading each simulation line with an intensity reflecting its centrality among the suite of simulations. This would mean that the more extreme simulations would be very pale and probably quite separated from each other, while the central simulations would be close together and darker.

Recommendation 17: Restricting the graph to the central 95% is a commonly used and entirely appropriate approach. However, consideration could also be given to other approaches that are designed to achieve a similar degree of interpretability.

Forecast and actual generation count and capacity tables

For example, consider Table 22 on page 64 of the “Forecast Accuracy Report 2022”, but these comments apply to all the Annual Reports and all the regions. The heading of the fourth block of columns in 2021 and 2022 has correctly been rewritten to follow the framing of comparing forecast to accuracy and states “(forecast – actual)”, while the column heading in 2020 was written as “(actual – forecast”). However, the calculation in these columns remains throughout the 3-year period to be of the form “(actual – forecast”).

Further to be consistent with the framing of comparing forecast to actual, these tables should be restructured so that the forecasts are provided to the left of the actuals.

Recommendation 18: Forecast and actual generation count and capacity tables should be restructured so that forecasts are provided to the left of the actuals and that the comparison column is calculated as (forecast – actual) so that the final column follows the generic definition of percentage error.

Regions providing generation support to other regions

In each Annual Report, it is stated that Tasmania is a winter-peaking region and that the availability of surplus generation provides important support to the mainland during summer peak demand events. In the 2020 Annual Report, it is mentioned that Queensland provides generation support to New South Wales during high-demand periods in New South Wales. These are examples of one region providing generation support to another region. It would therefore be insightful to provide an equivalent analysis of the accuracy of supply forecasts in the supporting region based on the top 10 hottest days in the supported region (or regions). For example, the report could present an equivalent analysis of the accuracy of supply forecasts in Tasmania based on the top 10 hottest days in Victoria (or perhaps across the rest of the NEM).

Recommendation 19: Consider providing an equivalent analysis of the accuracy of supply forecasts in the most important supporting regions based on the top 10 hottest days in the supported region.

Large-scale solar

We understand that the accuracy of Variable Renewable Energy (VRE) generation forecasts is presented by comparing modelled availability to actual generation, while all other generation types are presented by comparing modelled availability to actual PASA availability. In general, the difference between the two for VRE is assumed to be minimal. However, on two occasions (Victoria in 2022 on page 80 and Queensland in 2021 on page 65) significant excursions from this are noted and explained as being due to constraints representing system security and network limitations. Without more information, it appears that risks of such constraints might be ongoing and so could need further consideration in the modelling and reporting process.

We understand that VRE generation may be forecast both on an availability and a generation basis. To avoid the above issue confusing the accuracy of the forecasts, it would be worth investigating whether the accuracy of VRE generation forecasts could be presented consistently on a generation basis through both the forecast and reported data.

Recommendation 20: Consider whether the accuracy of VRE generation forecasts could be presented consistently on a generation basis through both the forecast and reported data.

Demand side participation – price triggers

The definition of the trigger events is potentially confusing. For example, the event set ">$300/MWh" includes all triggers no matter how high the price is, not just those with relatively low prices. Therefore, it is really a mix of low, medium and high prices and hence its median DSP response is probably that corresponding to a medium price event.

The choice of categories is an example of a bias/variance trade-off since the use of more, narrower categories reduces bias in the DSP response but introduces more variance in the DSP response due to the presence of fewer observations in each category. The current approach minimises variance but at the expense of significant bias. We believe that a better trade-off should be achievable by careful selection of new categories.

From a modelling and interpretation point of view, it would be better to split the trigger events into disjoint categories eg ">=$300/MWh AND <$500/MWh". From a modelling point of view, this would make the forecasts much easier to use, as one could then apply the appropriate (disjoint) category. Currently, one must try to apply the category ">$300/MWh" for only low triggers (as otherwise you would use a higher trigger category) and are then necessarily using events associated with much higher triggers. From an interpretation point of view, with disjoint triggers you would expect certain behaviours; for example, it would be reasonable to expect the forecast/observed median to increase with the higher triggers, since this is the rational response to a higher price.

Of course, the categories are only useful if they individually contain sufficient events to achieve sufficiently small variance in the estimate of the DSP response, so this proposed change may require the use of fewer categories. The current data would suggest that this would not be a significant loss as there is regularly not much signal in either the forecast or the actual across the current trigger categories. A price model that could be interrogated to understand the likely distribution of trigger categories would be helpful in defining the new disjoint categories.

Recommendation 21: Consider modifying the trigger categories into disjoint categories (eg ">=$300/MWh AND <$500/MWh" for the lowest trigger category). Further, consider reducing the number of categories to ensure sufficient events in each (disjoint) category while maintaining signal and interpretability.

The presented figures compare the forecast median with the observed median. While this is a legitimate comparison, it is potentially not very informative. It would be more informative to compare the forecast distribution against the observed distribution. This could still be done in one figure using side-by-side box plots or violin plots, for example.

Recommendation 22: Consider reporting the accuracy of the demand side participation forecasts by comparing the forecast distribution with the observed distribution, this could be achieved using side-by-side box plots or violin plots, for example.

Demand side participation – during reliability events

This is an interesting section and potentially one of significance. However, the presentation across the "Forecast Accuracy Reporting Methodology" paper dated August 2020 and the Annual Reports is a little confusing. In the Methodology paper, the description is based on regional maximum demand days and makes no mention of reliability events. In the Annual Reports, the title is "during reliability events" and the table reports forecasts during LOR2 and LOR3, but the discussions are focussed around maximum demand days.

It seems that there is a (perfectly reasonable) underlying assumption that maximum demand, high prices and reliability events are highly correlated. Under this assumption, it would be reasonable to conflate these ideas into the one concept. However, the discussions in the Annual Reports seem to suggest that this correlation may not be as high as one might expect.

Clarity on the purpose and meaning of this section would be useful. An investigation into the correlation between maximum demand, high prices and reliability events would be informative and might provide more direction to both the modelling of demand side participation and reporting of the accuracy of the forecasts.

Recommendation 23: Investigate ways to clarify the presentation in the "DSP response during reliability events" section to assist the reader's understanding.

6. Potential Modelling Improvements

In this section, we present some recommendations that are not directed at the annual accuracy reporting methodology itself but instead are suggestions that might improve the forecasting methodologies. It is quite possible that such changes would then lead to improvements in the annual accuracy reporting, but that is not the primary purpose of these recommendations. Two of the recommendations that have already been reported also fit partially into this category, namely Recommendation 6: and Recommendation 21:.

Analysis of consumption by industry segments

The introduction of refined industry segmentation for consumption forecasting is likely to lead to real improvements in the forecast and assessment of its accuracy. Investments in this area now will position AEMO well for the future as the economy transitions more rapidly to a low-carbon economy.

Recommendation 24: Consider the introduction of further industry segmentation to improve consumption forecasting.

More weather years would improve the robustness of forecasts

All the reports use weather years from the 2010-11 financial year onwards in the forecasts. This is a relatively small set and means that the forecasts can only consider a relatively narrow collection of weather possibilities. This necessarily has an impact on the potential accuracy of the forecasts through underestimating variance.

The Bureau of Meteorology has excellent weather data for the past 50+ years (at certain locations). We understand that there may be concern about using weather years from further back in history due to the effects of climate change introducing significant bias.

Hence, the choice of the range of weather years is a variance/bias trade-off.

It is possible to produce synthetic years that will allow the models to explore a broader collection of possible weather outcomes. This provides a more accurate representation of variance without introducing significant bias.

We briefly describe a possible way to generate synthetic weather years at existing Bureau of Meteorology sites. Synthetic weather years could be based on weather data taken from the global/regional climate models from the 2010-11 financial year through to 10 years ahead of the forecast (being the horizon of that forecast). For these to be comparable to the existing observation data, they would need to be statistically downscaled to the same sites. The process of statistical downscaling would be trained on the recent historical data (say 1980 – 2009) and then run on the period 2010 – 2030 (say). This process ensures that the results aren’t exposed to being biased by the climate model itself (some climate models are known to be hot, wet, hot and dry, etc) but instead provide many weather years that are compatible with the single realisation that we happened to observe.

Recommendation 25: To increase the robustness of the forecast process, increase the number of weather years that are incorporated. This could be achieved by using more historical data or, preferably, by making use of synthetic weather years.

Potential adjustment – voluntary load reductions

This feature plays only a small role in the forecasts, having been reported only once in the most recent three years across the five regions. Nonetheless, the assumed scaling factors are point estimates with no information regarding their accuracy or precision. If this feature is regarded as being of reasonable importance, then it is appropriate to investigate the accuracy of the assumed scale factors and possibly account for variability.

Recommendation 26: Consider upgrading the accuracy and level of assurance of the assumed scale factors in the Potential adjustment – voluntary load reductions feature.