THE UNIVERSITY *of* ADELAIDE

REVIEW OF FORECAST ACCURACY METRICS

For the Australian Energy Market Operator (2019)

adelaide.edu.au

CRICOS Provider Number 00123M

Citation: Cope, R.C., Nguyen, G.T., Bean, N.G., Ross, J.V. (2019) Review of forecast accuracy metrics for the Australian Energy Market Operator. The University of Adelaide, Australia.

Disclaimer

Purpose

The content of this report is provided for information purposes only.

No Warranty

Use of any content of this report is at your risk. Neither AEMO nor the University of Adelaide warrants or represents that the content of this report is complete or current or that it is suitable for particular purposes. You should verify and check the accuracy, completeness, reliability and suitability of any content from this report for any use to which you intend to put it and seek independent expert advice before using it.

Limitation of Liability

To the extent permitted by law, AEMO and the University of Adelaide and their advisers, consultants and other contributors to this report (or their respective associated companies, businesses, partners, directors, officers or employees) shall not be liable for any errors, omissions, defects or misrepresentations in the content of this report, or for any loss or damage suffered by persons who use or rely on such content (including by reason of negligence, negligent misstatement or otherwise). If any law prohibits the exclusion of such liability, AEMO's and the University of Adelaide's liability is limited, at their option, to the re-supply of the content, provided that this limitation is permitted by law and is fair and reasonable.

List of figures

Executive Summary

The Australian Energy Market Operator (AEMO) produces forecasts of *annual electricity consumption* and of *minimum/maximum half-hourly demand*, and must report at least annually on the accuracy of these forecasts. The University of Adelaide, School of Mathematical Sciences team were engaged to provide expert advice on the metrics used to assess forecast accuracy, as presented in the 2018 Forecast Accuracy Report (FAR) and in the internal performance monitoring dashboard (PD).

Broadly, current AEMO practices are appropriate and well-supported. We provide 14 recommendations, which are summarised on the following page, including both recommendations to continue current practice, and for improvements to forecast accuracy reporting and monitoring.

Forecasts of annual consumption consist of a point forecast of *annual operational consumption (sent out)* accompanied by point forecasts of various input drivers. AEMO's forecast assessments follow best practice and should continue in its current form (Rec. 1). Our two subsequent recommendations here (Rec. 2, 3) pertain only to communication of results, to provide additional context around the impact of input drivers.

Forecasts of seasonal minimum/maximum half-hourly demand are probabilistic, summarised in the FAR by reporting 10%, 50%, and 90% Probability of Exceedance (POE) forecasts. Forecast assessment is difficult as only one seasonal minimum/maximum demand observation occurs each year. This challenge is further exacerbated by the need to communicate forecast accuracy results across non-technical audiences. AEMO currently produces qualitative analyses and summaries of the drivers of minimum/maximum demand for the 2018 FAR and the Summer 2019 Forecast Accuracy Update; these should be continued (Rec. 4), with one recommendation on the communication of these results (Rec. 5).

Internally AEMO uses a range of more technical metrics to assess the accuracy of minimum/maximum probabilistic demand forecasts. Broadly, these are standard techniques for probabilistic forecast assessment, and are applied appropriately by AEMO. Specifically, for assessing probabilistic minimum/maximum demand forecasts AEMO consider both standard metrics for comparing distributions (the Mean Absolute Exceedance Probability and the Kolmogorov-Smirnov statistic) and for comparing competing forecasts (scores based on pinball loss). However, given the sparsity of available data, to construct these metrics it is necessary to produce more observations; one possible approach to this is to assess minimum/maximum demand forecasts over smaller time intervals (e.g., monthly). We recommend that the assumptions underlying this approach be carefully analysed to avoid introducing bias to the forecast assessment process (Rec. 6), and propose continued use with small modifications to these existing metrics (Rec. 7, 11). A *backcasting* approach was used in the 2018 FAR. We recommend that this be discontinued (Rec. 8); it appears that AEMO has independently done so, as this approach is not present in the 2019 summer FAR update. We also recommend that backcasting be replaced with a *full-season hindcasting* approach (Rec. 9), and that historical simulations also continue to be used as part of forecast assessment (Rec. 10).

Furthermore, we recommend that the distributions of residuals, currently used and assessed as part of the forecast development process, be incorporated more formally into the forecast assessment process through the PD (Rec. 12, 13), or similar dashboard. If the methodology used to produce probabilistic forecasts changes, these metrics should be assessed for relevance and replaced if required (Rec. 14).

List of Recommendations

This section contains a brief list of all (14) recommendations; further detail, justification and examples supporting each recommendation are presented throughout the report.

Annual consumption forecasts

Recommendation 1. Percentage error is a standard, useful and easily understood metric for comparing a point forecast to a point observation. It should continue to be used as the primary mechanism for assessing point forecast accuracy.

Recommendation 2. Wherever percentage error is reported for an input quantity, also indicate how the error of that input impacts the headline figure, *annual operational consumption (sent out)*.

Recommendation 3. Present the overall error in *annual operational consumption (sent out*) graphically as the sum of errors in input components, in a waterfall-plot.

Minimum and maximum probabilistic demand forecasts

Recommendation 4. The qualitative comparison provides clear and useful context around forecast accuracy. It should continue to be reported in the FAR.

Recommendation 5. Report probabilistic drivers of minimum/maximum demand graphically, overlaid with the actual value of the driver at the minimum/maximum demand interval.

Recommendation 6. Assess data aggregation processes to ensure that distributional assumptions are met. In particular, ensure that aggregation occurs on a scale that is relevant to business needs (i.e., seasonal minimum/maximum demand).

Recommendation 7. Empirical forecast distribution fit should continue to be assessed, using metrics such as the MAEP and Kolmogorov-Smirnov statistic. Care should be taken to ensure all distributional assumptions are met (following Rec. 6).

Recommendation 8. Normalise relative score by dividing the loss function by the true (observed) value, rather than the forecast quantile, so that the relative score is unbiased. The relative score should then be interpreted by taking the average of the loss function over all quantiles, with smaller values indicating better forecasts.

Recommendation 9. Discontinue backcasting as presented in the 2018 FAR. Replace with full season hindcasting (see Rec. 10 for details).

Recommendation 10. Perform *full-season hindcasting*: compare the forecast distribution that was made prior to a season (e.g., ESOO 2018 forecast), to the forecast distribution that would be made now using known inputs. This is to assess the impact of actual inputs on the forecast distribution produced.

Recommendation 11. Employ the *simulated history* approach: apply the current forecasting method to historical seasons, and compare against the observed minimum/maximum demand in those seasons. This provides more data with which to assess model accuracy and construct statistics such as the MAEP and KS statistic.

Recommendation 12. Analyse the observed residuals near the extremes of fitted demand, to ensure distributional assumptions made when forecasting are met.

Recommendation 13. Compare the residuals that produced the simulated seasonal minimum/maximum demands, to the observed residuals from the actual minimum/maximum demand intervals. This is to assess the plausibility of forecasting that observed minimum/maximum demands.

Recommendation 14. The strategies proposed in Rec. 12 and 13 are appropriate for the existing regression-simulation forecasting framework; if or when the forecasting methodology changes, these methods should be assessed for relevance and replaced with other (model-specific) diagnostics if necessary.

1. Introduction

AEMO produces forecasts of annual consumption and of minimum/maximum half-hourly demand. These forecasts are subject to high levels of scrutiny by various stakeholders, due to their importance in assessing supply adequacy, and the recent Retailer Reliability Obligation (RRO).

Forecast accuracy is reported no less than annually (as per clause 3.13.3A (h) of the National Electricity Rules) in the *Forecast Accuracy Report* (FAR). The purpose of the FAR is to build confidence in AEMO forecasts and to help inform the continuous improvement of those forecasts. The FAR has a wide audience including federal and state governments, industry forecasting practitioners, and the general public; as such it must present metrics that are appropriate and accessible. In addition, AEMO operates an internal *Performance Monitoring Dashboard* (PD), comprising more detailed and technical metrics of forecast accuracy and is directed towards a narrower, statistically-literate audience.

This report assesses the forecast accuracy metrics that are presented both in the FAR and in the PD. Note that this assessment applies to the final version of the 2018 FAR; subsequent modifications are outside the scope of this review, although some changes in the 2019 update are reported for context.

The report is organised as follows. Forecast assessment metrics for annual consumption are evaluated first, with associated recommendations provided. Then, forecast assessment for probabilistic minimum/maximum demand is evaluated, with each metric analysed and recommendations made in turn. Finally, additional forecast assessment methods for probabilistic demand are recommended.

The recommendations presented herein were developed based on publicly available reports, and some information around current PD presentation made available to the University of Adelaide team. As such, *these recommendations are conceptual in nature and will require internal AEMO assessment for technical feasibility*, based on internal expertise around the forecasting process, data availability and timeliness, etc. Specific details around data to be presented or prioritised should be decided internally, in collaboration with relevant stakeholders.

Note that examples presented herein are illustrative only, using synthetic data and simplified forecasting methods; they do not reflect actual AEMO data or forecast accuracy.

2. Annual Consumption

AEMO produces point forecasts of annual consumption, including both the headline forecast (i.e., *annual operational consumption (sent out)*) and a range of inputs that contribute to this headline forecast. Annual consumption consists of *residential consumption* and *business/industrial consumption*. The forecast of residential consumption is driven by consumer behaviour and normalised for annual variation in weather. The forecast of business/industrial consumption is based on a combination of surveys and econometric modelling. Input forecasts may comprise of a range of alternative scenarios.

2.1 Review of current approach to forecast assessment

Annual consumption forecasts are assessed on *annual percentage error* (PE), alongside visualisations of historical trends in PE and in overall annual consumption:

$$\mathrm{PE} = \frac{\mathrm{actual} - \mathrm{forecast}}{\mathrm{actual}} \times 100\ \%.$$

The raw magnitude of the error is also reported.

A perfect forecast would produce a PE of 0%. Historical values of PE for annual operational consumption (sent out) have generally been low (e.g., within $\pm 10\%$). Individual input forecasts often have higher errors.

Communication: In the 2018 FAR, annual consumption forecast accuracy is reported first as an overall summary of PE by state and aggregate (Table 4 2018 FAR; including brief context explaining the possible sources of error in each state). This is accompanied by summaries of three key drivers. Subsequently, each state is detailed individually (e.g., Table 8 2018 FAR, NSW). This included a range of related generation/consumption metrics, some significant input forecasts, and weather factors (i.e., the number of heating and cooling degree days that occurred). The significant input forecasts reported in the 2018 FAR are *transmission losses* and *rooftop PV generation offset* in every state, with the addition of *coal seam gas* in Queensland.

In addition, the PD includes figures of trends of PE over time for each scenario, when inputs are scenario-based. This provides additional context around how effectively the different scenarios have reflected the actual driving inputs over time.

2.2. Recommendations

Recommendation 1. Percentage error is a standard, useful and easily understood metric for comparing a point forecast to a point observation. It should continue to be used as the primary mechanism for assessing point forecast accuracy.

As this forecast accuracy metric for annual consumption is appropriate in its current state, the following two recommendations pertain solely to the presentation of results in the FAR.

Recommendation 2. Wherever percentage error is reported for an input quantity, also indicate how the error of that input impacts the headline figure, *annual operational consumption (sent out)*.

For example, consider reporting of the key input *Gross State Product* (GSP). Compute the difference between: (i) the *annual operational consumption (sent out)* with the actual value of GSP (and all other known inputs), and (ii) the *annual operational consumption (sent out)* with the forecast value of GSP (with all other known inputs). Report this difference as a percentage of (actual) *annual operational consumption (sent out)*.

This approach is of value as some inputs have relatively large PE, but those errors may have a relatively small impact on overall error in *annual operational consumption (sent out)*. Quantifying their indicative impact provides additional context and may provide guidance as to which inputs should be prioritised to improve future forecasts.

This approach should be applied to all possible inputs, including weather. It is applicable to both the *key input drivers* reported at the national level (Section 2.2, Tables 5-7 2018 FAR), and to the detailed state-by-state analyses of annual consumption. An example of the proposed modification to the state-by-state results appears in Fig. 1. Note that some actuals are model estimates, which may change year-on-year. Further, some components may not be observed or reported immediately. The information presented should be selected to ensure feasibility, guided by AEMO expertise on how each component is observed/estimated, data availability, and with stakeholder input.

Annual consumption	**2017 ESOO forecast**	**Actual**	**Difference**	**Difference (%)**	**Indicative impact on `sent out` consumption %**
Operational consumption – sent out (GWh)	67,819	67,899	80	0.1%	0.12%
Generation forecasts					
Auxiliary load (GWh)	3,996	3,105	-891	-28.7%	-1.31%
Operational consumption – as generated (GWh)	71,815	71,004	-811	-1.1%	-1.19%
Non-scheduled generation (GWh)	1,652	2,070	418	20.2%	0.62%
Native consumption – as generated (GWh)	73,467	73,074	-393	-0.5%	-0.58%
Significant input forecasts					
Transmission losses (GWh)	872	1,556	684	44.0%	1.01%
Rooftop PV generation offset (GWh)	-1,991	-2,068	-77	3.7%	-0.11%
Weather factors – annual					
Heating degree days (HDD)	618	640	22	3.4%	0.32%
Cooling degree days (CDD)	449	577	128	22.2%	1.09%

Figure 1. Example recommended modification of state-based annual consumption reporting. Proposed additional column is bordered in red. Modified from Table 8, (NSW) 2018 FAR. Note that indicative impact for weather factors presented here are not actual values.

Recommendation 3. Present overall error in *annual operational consumption (sent out)* graphically as the sum of errors in input components, in a waterfall-plot.

The error in *annual operational consumption (sent out)* is driven by the errors in the various input forecasts; understanding the contribution of each of those inputs to the overall consumption would be of value. A waterfall plot is ideal for visualising cumulative errors. Inputs should be organised in an informative way, as appropriate, e.g., by grouping errors in *Transmission*, *Residential*, *Industrial*, and *Business consumption*. Fig. 2 presents an example of this type of visualisation. The grouping and presentation of components should be determined based on internal AEMO expertise and in collaboration with stakeholders.

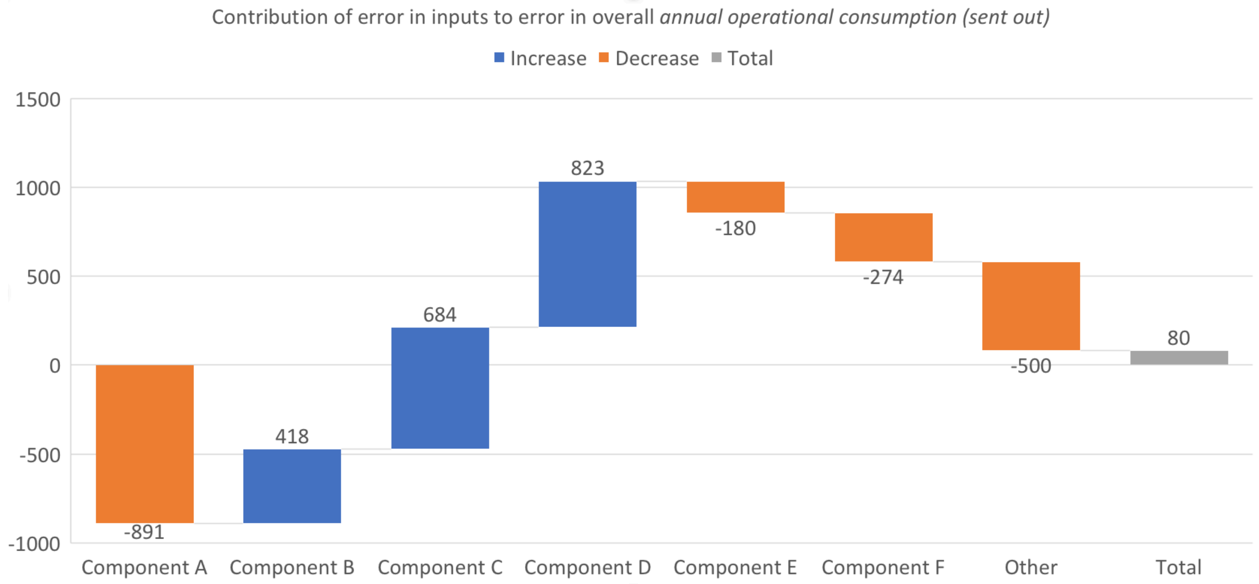


Figure 2. Example waterfall plot showing the contribution of individual input errors to the overall error in *annual operational consumption (sent out)*. The inputs to be presented should be selected in collaboration with stakeholders.

3. Minimum and maximum probabilistic demand

AEMO produces forecasts of the distributions of seasonal minimum and maximum half-hourly demand. These forecasts have historically been calculated through a computationally-intensive simulation process[1]:

- building synthetic temperature-years by bootstrapping two-week intervals of historical temperature data (modified to account for climate change);
- projecting (half-hourly) regression models relating demand to temperature onto synthetic temperature data, including stochastic volatility (Gaussian errors); and,
- extracting the seasonal minimum/maximum demand.

However, AEMO may develop new methods for forecasting minimum and maximum probabilistic demand going forward, and so any forecasting accuracy metric that depends on the forecasting process might need revision at that point.

The distributions of possible seasonal minimum or maximum demand are summarised in the FAR by reporting 10%, 50%, and 90% Probability of Exceedance (POE) forecasts (see e.g. Fig. 3). The 2018 FAR (p.10-16) presents substantial detail on how these forecasts were produced.

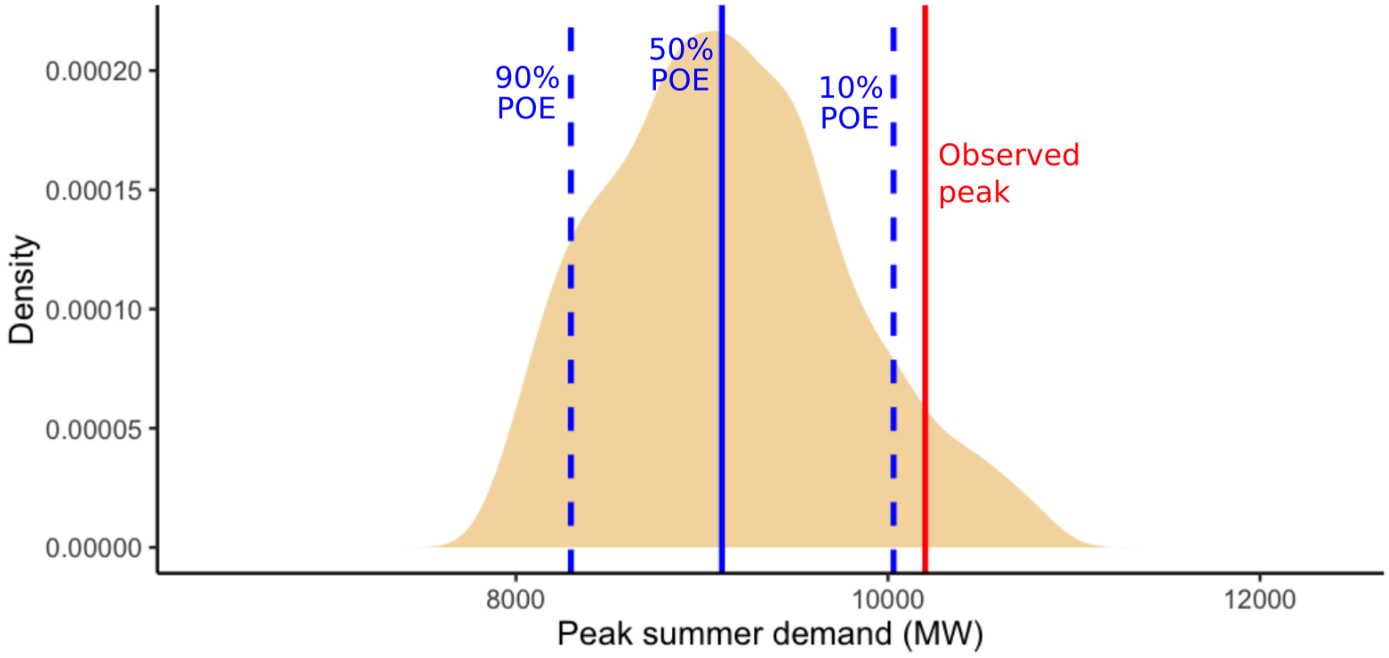


Figure 3. Visualisation of a probabilistic forecast of maximum demand, the reported POE levels, and a hypothetical observation of maximum demand from that season. This illustrates the difficulty of forecast assessment with only one observation: an observation above the 10% POE should be rare, but is still consistent with the forecast.

Note that the simulations that generate these forecast distributions are not recorded for further analysis. We produce example data from a similar process (with fewer inputs) for illustrative purposes.

[1] See Electricity Demand Forecasting Methodology Information Paper (Draft updates for 2019 Electricity Statement of Opportunities), April 2019, AEMO.

Overview of current approaches to probabilistic forecast assessment

It is challenging to retrospectively assess the accuracy of a probabilistic forecast against a single point observation. Communicating forecast accuracy across a range of stakeholders provides a further challenge, given the highly technical nature of probabilistic forecast assessment. No single forecast assessment method is sufficient for assessing these probabilistic forecasts; as such, AEMO produces a range of different methods, each of which provides different insights.

Each current probabilistic forecast assessment metric is assessed in detail below. The 2018 FAR presented:

- a qualitative comparison of the observed minimum/maximum demand to the forecast distribution;
- a distribution for individual key drivers (e.g., temperature, time of day); and,
- a *backcast* of the top 15 observed demand intervals.

The PD includes more technical approaches, such as:

- use of the Kolmogorov-Smirnov statistic and the Mean Absolute Excess Probability to compare forecast and observed distributions;
- relative scores based on the pinball loss function; and,
- simulated histories.

Some of these more technical approaches appeared in the 2015 FAR but were removed from subsequent reports due to stakeholder feedback. The 2019 summer FAR update excluded the backcast of the top 15 observed demand intervals, and presented further qualitative analysis of the assessment of maximum demand.

3.1 Qualitative comparison [FAR]

The primary method of reporting accuracy of minimum/maximum demand forecasts in the FAR is qualitative comparison: specifying where on the forecast distribution the observed minimum/maximum demand lies, and providing contextual factors that may explain this.

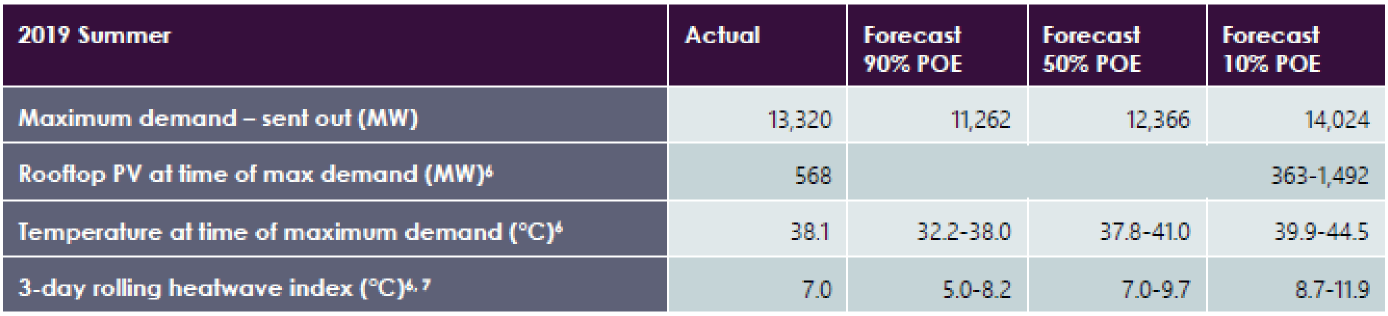
An example from the 2017 FAR is:

> "In NSW 2017, Maximum demand occurred on 10 February 2017, when the temperature reached 43.7°C. The actual MD may have been higher if it hadn't been for a general call for reduced consumption and engage DSP. Accounting for an estimated combined 490 MW of load reductions, the adjusted MD would have exceeded the forecast of 10% POE demand."

Recommendation 4. The qualitative comparison provides clear and useful context around forecast accuracy. It should continue to be reported in the FAR.

3.2 Probabilistic distribution of drivers of minimum/maximum demand [FAR]

Alongside the forecast distribution of minimum/maximum demand, the distribution of key drivers at the (simulated) minimum/maximum demand(s) are also reported. Specifically, in the 2019 Summer FAR update a range for each driver corresponding to each reported POE level is included in a table, e.g., as in Fig. 4. In the 2018 FAR these were reported as single values rather than ranges.



2019 Summer	Actual	Forecast 90% POE	Forecast 50% POE	Forecast 10% POE
Maximum demand – sent out (MW)	13,320	11,262	12,366	14,024
Rooftop PV at time of max demand (MW)[6]	568	363-1,492		
Temperature at time of maximum demand (°C)[6]	38.1	32.2-38.0	37.8-41.0	39.9-44.5
3-day rolling heatwave index (°C)[6, 7]	7.0	5.0-8.2	7.0-9.7	8.7-11.9

Figure 4. Example of current presentation of probabilistic distribution of drivers. From Table 5, Summer 2019 FAR Update.

Reporting these quantities at the observed minimum/maximum demand interval, compared to the distribution that produced the forecast, provides valuable context. However, the clarity of these results could be improved by displaying histograms of these drivers rather than as summary values or ranges in a table.

Recommendation 5. Report probabilistic drivers of minimum/maximum demand graphically, overlaid with the actual value of the input at the minimum/maximum demand interval.

Fig. 5 provides an example of this approach. Visualising the full distribution, instead of providing a three-point summary, provides a more concrete understanding of the distribution of drivers. For space efficiency, it may be necessary to present some of these figures in an appendix or in an online supplement rather than the main report. The PD could also be used to explore bivariate distributions of input drivers.

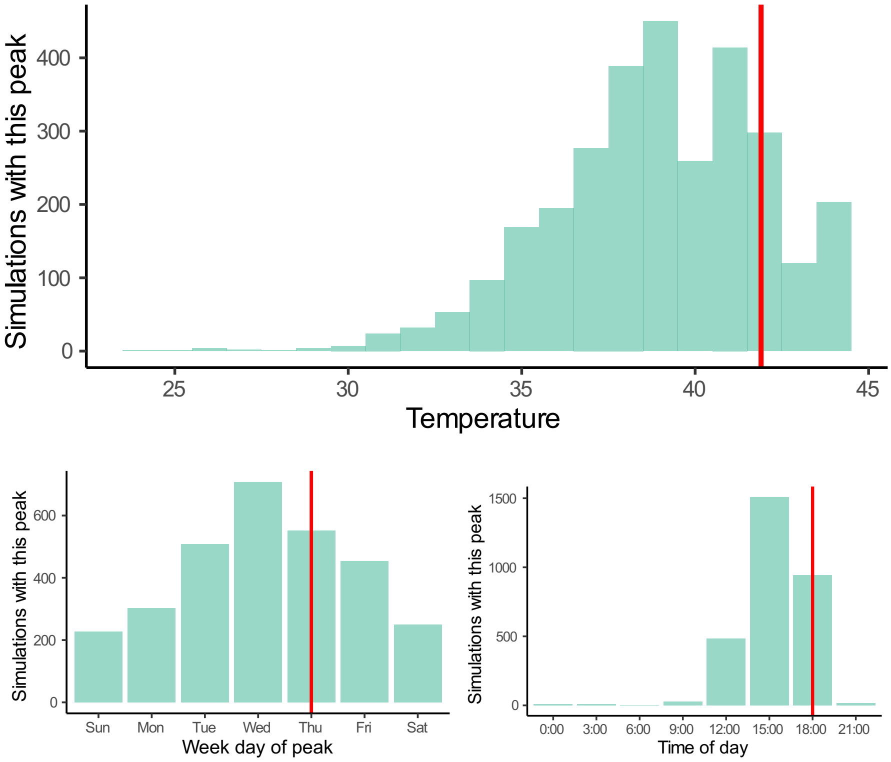


Figure 5. Example visualisation of probabilistic drivers of demand. The histogram shows the distribution of the driver at the forecast, the red vertical line the value of that driver at the true minimum/maximum demand interval. Note: not an actual AEMO forecast.

3.3 Data aggregation

To effectively assess a probabilistic forecast, many (independent) observations from that distribution are required. For example, a 10% POE forecast for maximum demand means that the seasonal maximum should exceed that POE 10% of the time; to test this, if one could produce forecasts of the seasonal maximum over many years, 10% of the observed values should exceed that POE. This is the *Law of Large Numbers*. If, instead, the proportion exceeding the POE differed substantially from 10%, that may be statistical evidence that the forecast is inaccurate. How strong that statistical evidence is, depends on both the number of observations and the accuracy of the forecast (highly inaccurate forecasts will be easier to detect). Appendix B provides an example of this relationship.

Consequently, effectively producing more data on which to assess forecasts is critical to forecast accuracy assessment. Possible approaches include aggregating across states, or disaggregating over time. For example, producing forecasts of monthly or weekly minimum/maximum demand, and using the assessment of those disaggregated forecasts to assess seasonal forecast accuracy. However, this approach may violate the distributional assumptions necessary to assess forecast accuracy. First, weekly or monthly forecasts in the same location are unlikely to be independent. For example, a heatwave at the end of one month may continue to the start of the subsequent month; more broadly, months within the same season will be subject to the same model conditions (e.g., El Niño or La Niña), giving the impression of systematic bias in forecasts. Second, the minimum/maximum demand over shorter time intervals (weeks or months) might not have the same distribution as the seasonal minimum/maximum demand. Inappropriate use of aggregated data could bias estimates of accuracy for the actual quantities of business need (i.e., seasonal minimum/maximum demand).

Recommendation 6. Assess data aggregation processes to ensure that distributional assumptions are met. In particular, ensure that aggregation occurs on a scale that is relevant to business needs (i.e., seasonal minimum/maximum demand).

3.4 Mean Absolute Excess Probability & Kolmogorov-Smirnov statistic [PD]

As the forecast of maximum or minimum demand is a distribution, standard tools for assessing observed empirical distributions against expected (i.e., forecast) distributions may be applied. If multiple observations are available (following Recommendation 6), these approaches can assess the degree to which the forecast distribution and the observed distribution correspond (across the full distribution). This approach was presented in the 2015 FAR, and is currently used in the PD.

Given a percentile p, $G(p)$ is the proportion of the empirical observations that exceeds the pth percentile of their forecast minimum/maximum demand distribution. A perfect forecast would approach $p = G(p)$, for all p, with sufficiently many observations, i.e., 90% of observations exceed the 90% POE, 50% exceed the 50% POE, 10% exceed the 10% POE, and so on. Fig. 6 demonstrates this relationship.

From this, two metrics of forecast accuracy are computed:

- the Kolmogorov-Smirnov (KS) statistic, $\mathrm{KS} = \max_{p} |G(p) - p|$; and,
- the Mean Absolute Excess Probability (MAEP), $\mathrm{MAEP} = \int_0^1 |G(p) - p| dp$.

These are standard metrics used to assess density forecasts[2], requiring no assumptions on the distribution of the extreme values. AEMO currently uses them as raw metrics of forecast accuracy --- to make comparisons between reports or across regions (e.g. Fig. 7) --- rather than as part of a statistical test. A perfect forecast would have both KS and MAEP approach 0 as the number of observations increases (i.e., to infinity). A standard test for the KS statistic exists and is available in any statistical computing package; however, this test requires a relatively large sample size, which is unlikely to be available to AEMO in the foreseeable future.

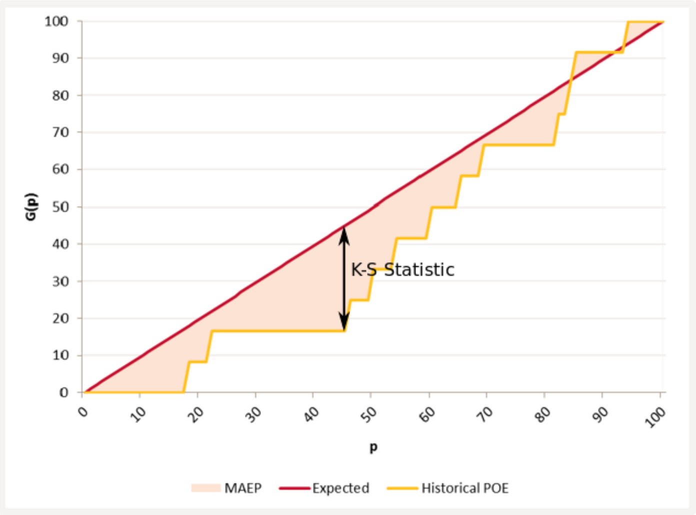


Figure 6. Visualisation of the KS Statistic and Mean Absolute Excess Probability (MAEP). Adapted from Fig. 2, 2015 FAR.

These two related metrics, and more generally visualisations comparing the distribution of observed POEs against the expected distribution, are key tools for forecast evaluation[3]. Specifically, they describe the overall consistency between the forecast distribution and the observations, penalising forecasts with consistently too little or too much variance. Moreover, they are on an easily interpretable scale which does not depend on the magnitude of the observations or the forecasts (MAEP is always between 0 and 0.5; the KS statistic is always between 0 and 1). This feature makes it possible to compare forecast accuracy between different regions, different years, and different seasons. The limitation of KS and MAEP is the need for many observations, and that it does not assess forecast sharpness (i.e., how concentrated the forecast probability mass is around the observed minimum/maximum demand) as effectively as other methods[4]. These metrics are suitable for the PD (as currently used), as the interactive nature of the dashboard allows for exploration of how variations on the data included in the empirical distribution impact the statistics.

Recommendation 7. Empirical forecast distribution fit should continue to be assessed, using metrics such as the MAEP and Kolmogorov-Smirnov statistic. Care should be taken to ensure all distributional assumptions are met (following Recommendation 6).

[2] Mitchell, J., & Wallis, K. F. (2011). Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness. *Journal of Applied Econometrics*, *26*(6), 1023-1040.

[3] Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2), 243-268.

[4] Hong, T., & Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, *32*(3), 914-938.

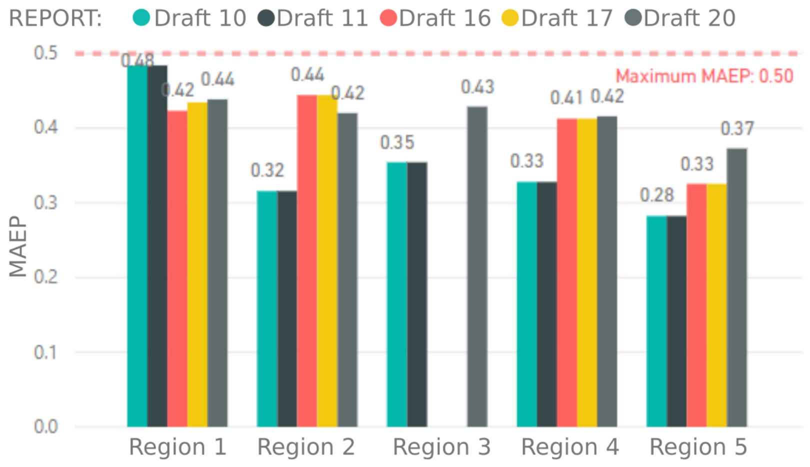


Figure 7. Example of how the metric MAEP is used to compare forecast accuracy across regions, and reported in the PD.

3.5 Score and Relative Score [PD]

The score is a metric that encapsulates both how close a probabilistic forecast distribution is to the target quantity, and how confidently it makes that prediction. It is the metric used to compare forecasts in a range of contexts, including the global energy forecasting competitions in 2014[5] and 2017[6]. However, the scale of the score is dependent on the data being forecast, and thus cannot be directly compared, e.g., across states or seasons. AEMO has modified the metric to produce the relative score, which is on a similar scale regardless of the data used. Broadly, we support the use of this metric, but suggest an alternative modification to produce a new relative score; details follow.

Given an observed maximum demand y, a quantile p, and corresponding forecast at that quantile q_p, the *pinball loss* function is

$$L(y, p, q_p) = \begin{cases} (1-p)(q_p - y) & y < q_p, \\ p\,(y - q_p) & y \geq q_p. \end{cases}$$

Intuitively, it represents the distance between the observation and the forecast at a quantile, weighted by the tail probability at that quantile. The average of this loss function is taken over all quantiles p to produce the percentile score. When multiple forecasts occur on the same scale, these scores may be meaningfully compared: smaller values indicate better forecasts.

This is an excellent metric for comparing forecasting methods given few observations of the target quantity, as it favours forecasts that are accurate, and *sharp* (i.e., the majority of forecast probability mass is near the observation).

However, the scale of the score depends upon the scale of the observations, and it does not have an intuitive interpretation. Consequently, the score is only meaningful when comparing between

[5] Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*.

[6] Hong, T., Xie, J., & Black, J. (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*.

forecasting methods on the same data. To address this issue, AEMO have developed the *relative score*, by normalising the pinball loss function by the forecast quantile:

$$\mathrm{RL}(y, p, q_p) = \begin{cases} \dfrac{p(y - q_p)}{q_p} & y > q_p, \\ \dfrac{(1-p)(q_p - y)}{q_p} & y \leq q_p. \end{cases}$$

The relative loss function is averaged to obtain the relative score, used to produce comparisons between reports and locations (Fig. 8), and a comparison across percentiles.

The relative score provides a way to compare years, locations, and models on a similar scale. For example, Fig. 8 indicates that Draft 20 forecast accuracy was consistently worse than other forecasts shown, and that the Draft 16 and Draft 17 forecasts are consistently very similar. This approach is particularly useful when it is necessary to assess accuracy of multiple years of forecasts simultaneously. This is due to it being straightforward to evaluate which data were generally easier, or more difficult, to forecast across candidate methods, or which methods were consistently able to produce the best forecasts.

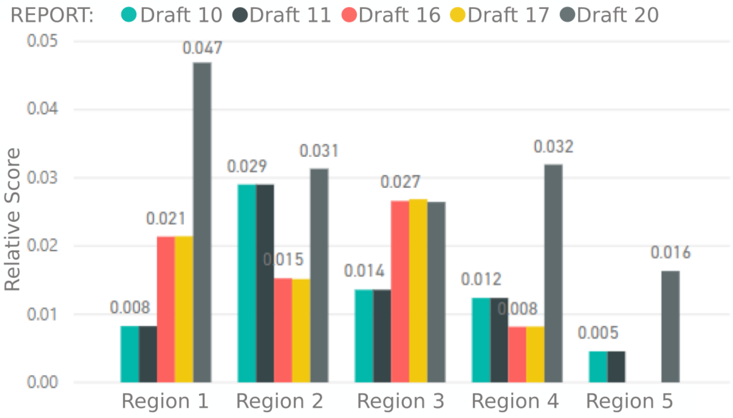


Figure 8. Example of current usage of relative score (by region and report, i.e., averaged over quantiles). From the PD.

However, scaling the score by the forecast at each percentile (q_p), rather than the actual value (y), produces a score that is not symmetric around the true value. Consequently, the relative score is biased: when the forecast underestimates the value, the relative score is higher than when the forecast overestimates the value by the same amount (see Appendix A).

Recommendation 8. Normalise relative score by dividing the loss function by the true (observed) value, rather than the forecast quantile, so that the relative score is unbiased. The relative score should then be interpreted by taking the average of the loss function over all quantiles, with smaller values indicating better forecasts.

The appropriate relative loss function is then:

$$\mathrm{RL}_{\mathrm{proposed}}(y, p, q_p) = \begin{cases} \dfrac{p(y - q_p)}{y} & y > q_p, \\ \dfrac{(1-p)(q_p - y)}{y} & y \leq q_p. \end{cases}$$

This is now symmetric and therefore unbiased, as illustrated in Fig. 9.

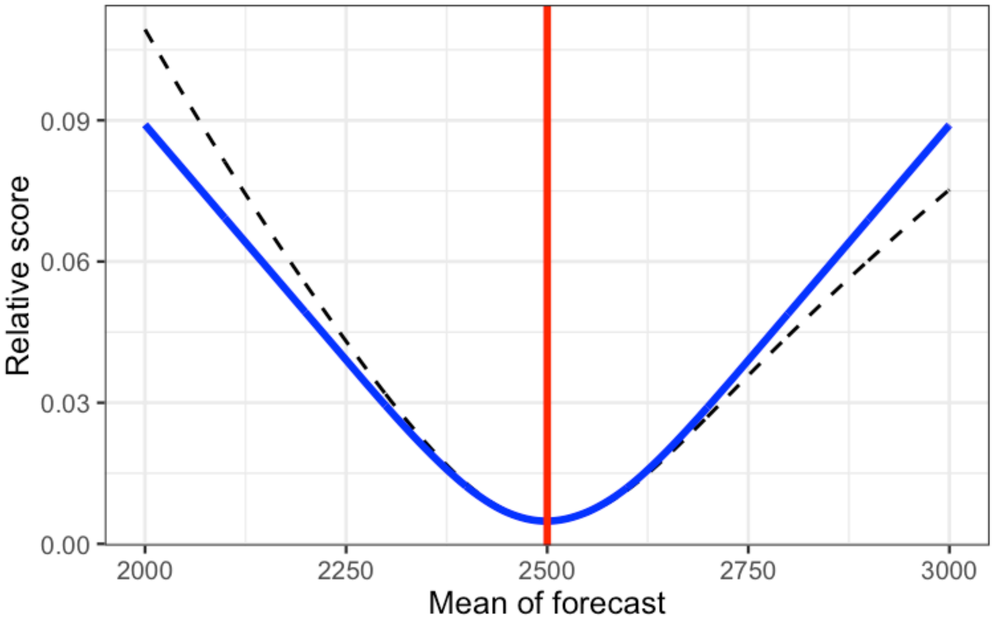


Figure 9. Relative score as a function of the mean of the forecast, under the existing (black) and proposed (blue) relative loss function.

Note that, normalising by the actual value rather than the forecast is consistent with the normalisation applied to annual consumption (i.e., to calculate percentage error).

Once this is resolved, the relative score metric is suitable for comparisons between methods given the same data. The proposed relative score metric effectively puts forecasts on a similar scale for assessment. We note that direct comparisons between locations or across different years of data should be made with caution; while relative scores are on the same scale, the score is designed to compare forecasts given the same data, not forecasts of different quantities. Furthermore, the relative score should be considered only when taking the average of the loss function over all quantiles; considering the loss function at fixed quantiles and averaged across locations is not informative.

3.6 Backcasting minimum/maximum demand events [FAR]

In the 2018 FAR, *backcasting* was performed by evaluating the forecasting regression (without stochastic volatility) at each of the 15 highest observed demand intervals in each region. Fig. 10 illustrates how this is presented.

This method demonstrates that the intervals with highest observed demand were (generally) those for which the observed value was greater than the mean of the forecast regression with those predictors (i.e., without stochastic volatility). In effect, these high values could be interpreted as having positive residuals. The backcasting approach does not provide information regarding the accuracy of the forecast. AEMO has clearly recognised this, as the approach is no longer present in the most recent summer 2019 FAR update.

Recommendation 9. Discontinue backcasting as presented in the 2018 FAR. Replace with full season hindcasting (see Rec. 10 for details).

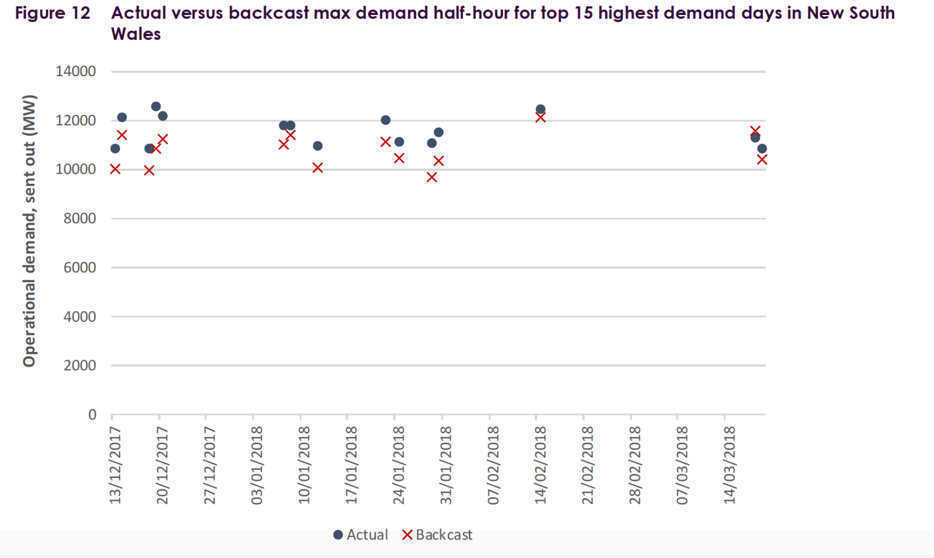


Figure 10. Example of current presentation of backcasting. From Fig. 12, 2018 FAR.

3.7 Hindcasting and simulated history

These two related and complementary methods, detailed below, use historical data and forecasts to inform forecast accuracy assessment.

- *Full-season hindcasting* involves comparing a forecast that was made historically (e.g., the ESOO 2018 forecast) to the forecast that would be made now for that year (i.e., with actuals for temperature, growth drivers, etc.). The purpose of this method is to compare the forecast distribution from a previous forecast to the forecast distribution that would be made now using the *actual* input drivers. The aim is to elucidate the impact of the forecast drivers (synthetic temperature years, growth drivers) on the forecast.
- *Simulated history* involves applying the current forecasting method to historical data to generate historical distributions of minimum/maximum demand. These are compared against the observed minimum/maximum demands in historical years for assessment. That is, the current forecasting method is applied to forecast maximum demand in each of 2017, 2016, and so on, and each is compared against the corresponding observed maximum demand in that year. This produces more data points that can be used to test the forecasting method itself, rather than the forecast that was produced in a given year.

Each of these is now expanded upon.

3.7.1 Full-season hindcasting [new method – FAR or PD]

To produce the probabilistic forecasts, regression models are fit to historical demand data, and then simulated 3,000 times on synthetic weather data (constructed from 20 years of historical weather) with stochastic volatility[7]. The minimum/maximum demand observations from the 3,000 simulations are extracted, to produce the probabilistic demand forecast distributions.

To hindcast, the same process is followed, but rather than using synthetic weather years (constructed from 20 years of historical data), the actual (full) year of temperature should be used. The process is repeated 3,000 times, with the whole year simulated in each case, and the half-hourly minimum/maximum demand in each simulation is extracted to produce the hindcast

[7] For more detail please see AEMO's Electricity Demand Forecasting Methodology Information Paper.

distribution. This allows hindcasting to capture the impact of model stochastic volatility, and other inputs (e.g., solar). We emphasise that hindcasting is a full-year process, and is not informed by the actual day on which minimum/maximum demand occurred.

Recommendation 10. Perform *full-season hindcasting*: compare the forecast distribution that was made prior to a season (e.g., ESOO 2018 forecast), to the forecast distribution that would be made now using known inputs. This is to assess the impact of actual inputs on the forecast distribution produced.

Comparing the forecast distribution with this hindcast distribution identifies the impact of climate, represented through the synthetic weather-years, on the process. For example, consider Fig. 11. In this scenario, the synthetic weather years impact the forecast substantially: they lead to underestimates of the maximum demand.

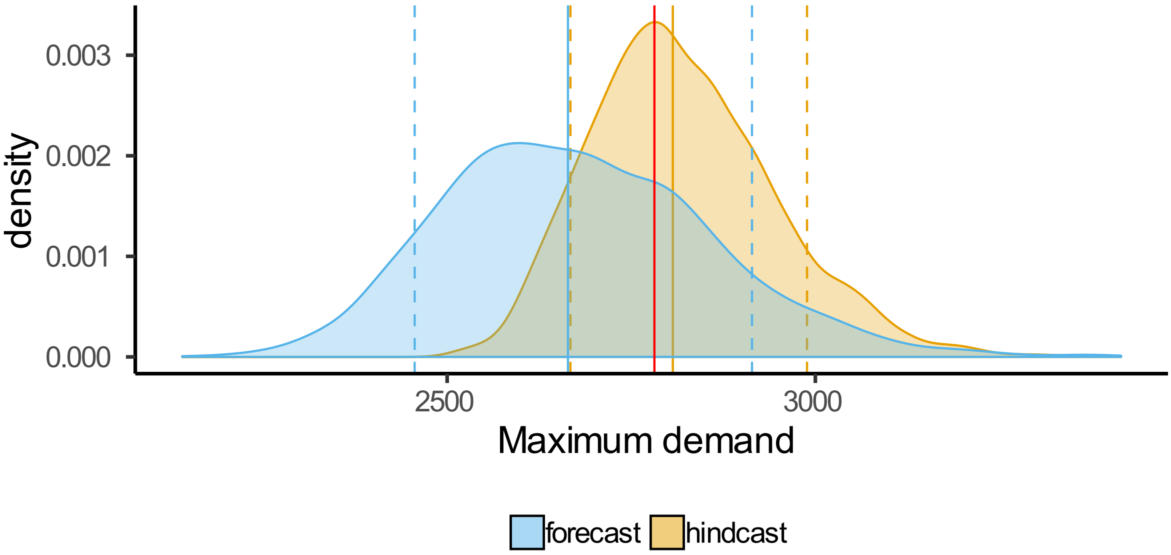


Figure 11. Comparison between forecast and hindcast distribution of maximum demand. Note: not an actual AEMO forecast.

A similar process could be applied with the other inputs, in order to elucidate the marginal impact of each input on the forecast distribution. (For example, one could use known (true) values of solar generation, but synthetic temperature, in order to quantify the impact of solar generation input on the forecast distribution.) A visualisation of this form may be appropriate for inclusion in the FAR. Note that we do not expect the hindcast distribution to be identical to the forecast, as they were constructed from different data, but differences between the distributions may provide insight into possible forecast errors.

Under this framework, the probabilistic distribution of input drivers between the forecast and the hindcast can be compared (e.g., Fig. 12). As weather data correspond to the actual day on which they occurred, it is also appropriate to compare the dates on which the hindcast produced minimum/maximum demand to the date of the actual minimum/maximum demand interval (Fig. 13). In this example, using actual weather data (i.e., in the hindcast) we would predict that the maximum summer demand would be most likely to have occurred during either the mid-January or early-February heatwaves. In fact, peak demand did occur on the most likely hindcast day in mid-January. These types of analyses may be appropriate for the PD. We emphasise that considering hindcasts of single days is not appropriate; rather the whole year of actual weather observations must be simulated to correctly produce the hindcast distribution.

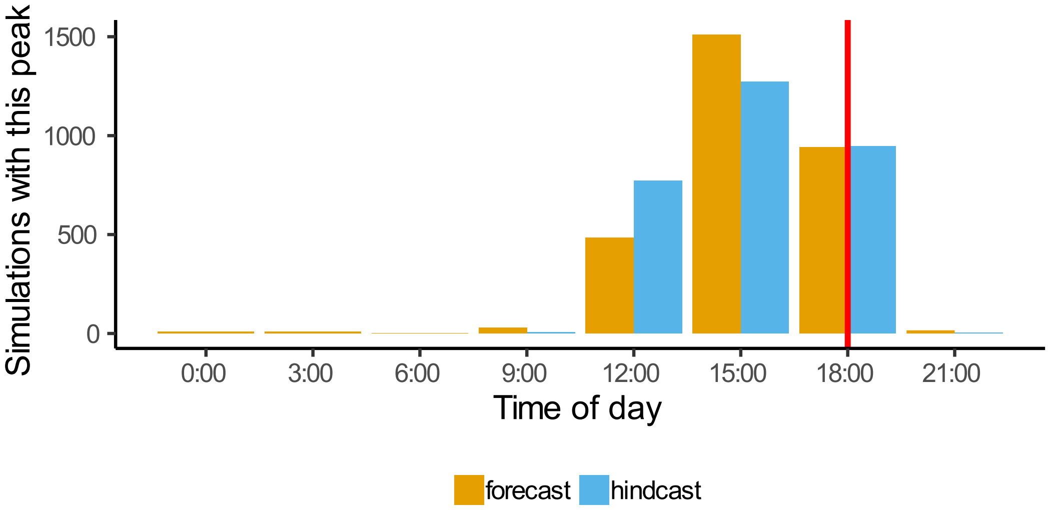


Figure 12. Comparison between time of day for maximum demand between forecast and hindcast distributions. Note: not an actual AEMO forecast.

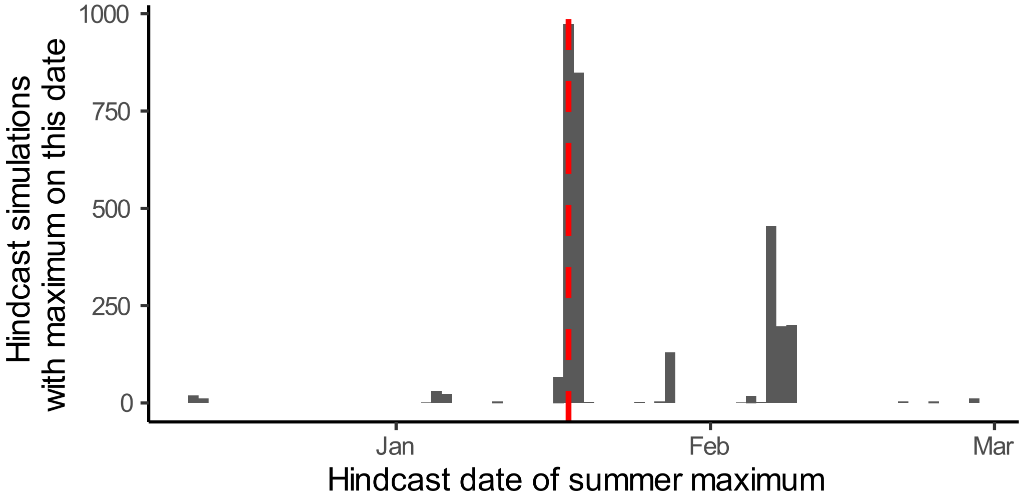


Figure 13. Hindcast summer maximum date of occurrence (December 2017—March 2018). Red line indicates the actual date of maximum demand. Note: not an actual AEMO forecast.

3.7.2 Simulated History [PD]

Forecast methodology changes each year, and so historical forecast accuracy results do not inform current forecast accuracy assessment. However, applying current methods to past seasons can provide some indication of how the correspondence between forecasts and actual minimum/maximum demand would have varied over time. As such, the *simulated history* approach is to:

- apply the current forecasting model to historical data from previous years;
- produce probabilistic forecasts of minimum/maximum demand for these years; and,
- compare the actual minimum/maximum demand to those historical 'forecast' distributions.

Fig. 14 shows an example output of this process: for each season, the simulated history forecast POE levels are compared to the actual minimum/maximum demand that occurred in that season. Then, if the forecast is accurate and we had sufficient data, we expect 10% of seasons to exceed the 10% POE, 50% of seasons to exceed the 50% POE, and so on. This approach was presented in the FAR in 2015, and is currently used in the PD.

We note that these must be interpreted in context. That is, conditions around demand might have changed over time and so it may not be the case that a forecast designed for 2018 is suited to forecasting demand in 2008 (e.g., due to changes in the operational environment); this should be assessed. When it is determined that simulated history can be applied effectively, these outputs could be used to calculate MAEP/KS statistics that avoid some of the limitations around monthly disaggregation, for example. Further, the number of observations available must be considered when assessing accuracy with respect to POEs. For example, given sufficiently many years of data we would expect 10% of seasonal maximums to exceed the 10% POE; however, with few observations (e.g., <10) there is insufficient statistical evidence to detect forecast errors in the majority of cases (see Appendix B).

Recommendation 11. Employ the *simulated history* approach: apply the current forecasting method to historical seasons, and compare against the observed minimum/maximum demand in those seasons. This provides more data with which to assess model accuracy and construct statistics such as the MAEP and KS statistic.

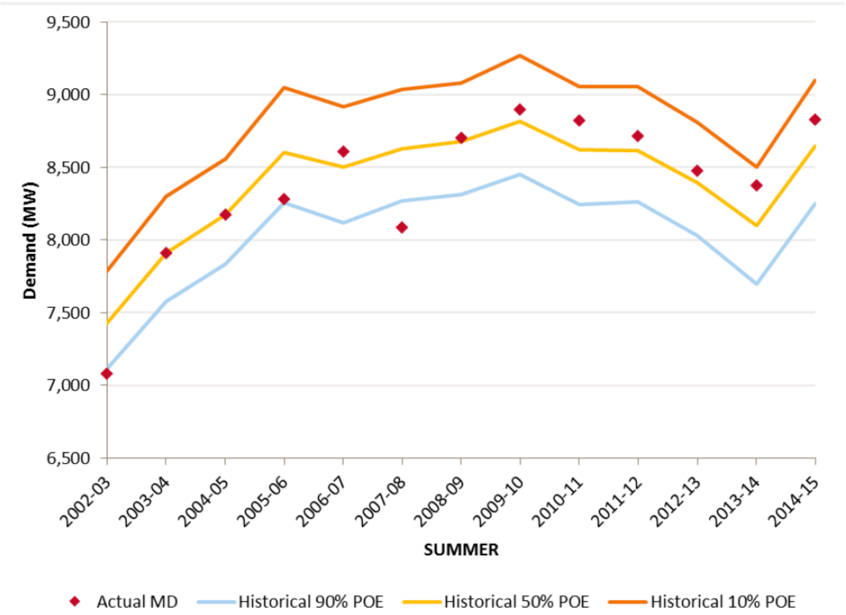


Figure 14. Example of visualisation of simulated history. From Fig. 6, 2015 FAR.

4. Proposed methods for probabilistic minimum/maximum demand forecast assessment

Because of the challenges of probabilistic forecast assessment when few data points are available, it is appropriate to consider forecast assessment methods which more explicitly test the assumptions of the forecasting method itself. Recommendations 12 and 13 are examples that may be suitable for this purpose: they focus on testing the model residuals, as these have a substantial impact on the minimum/maximum demand in the current forecasting methodology. Discussions with AEMO throughout the preparation of this report indicated that these methods are broadly similar to ideas currently used internally during forecast model development. If the forecasting method itself changes, these specific recommendations may no longer be applicable, and instead they should be replaced with metrics relevant to the new forecasting approach (Recommendation 14 formalises this).

Recommendation 12. Analyse the observed residuals near the extremes of fitted demand, to ensure distributional assumptions made when forecasting are met.

Each observation, y, in the year being forecast, has predictors, x. If the forecast model is $f(x) + \varepsilon$, the observed residual from that forecast is

$$e(x, y) = y - f(x).$$

The extreme values of $f(x)$ should be those that are most likely to produce extreme demands. Therefore, the residuals of these observations should be analysed.

In a good forecasting model, under the regression-simulation framework, the observed residuals $e(x, y)$ should have the same distribution as the errors ε used in the forecast simulations. To analyse the residuals of the extreme values, first, the minimum and maximum 5% of values by $f(x)$ should be extracted. Then, their residuals tested for normality (e.g., by the Shapiro-Wilk test) and for equal variance with the distribution of stochastic volatility used to produce the forecast. Fig. 15 and 16 demonstrate this process for a forecast that is not accurate; Fig. 15 shows that the residuals are heteroscedastic, and Fig. 16 shows that the distribution of the observed residuals does not match that which generated the forecast.

Assessing residuals versus fits in this way is a standard approach when building regression models, and is used during model development at AEMO. We recommend AEMO formally include this as part of its forecast assessment. This is critical to forecast assessment due to the way that forecasts are currently produced: if actual residuals from the forecast regression at the extremes do not match the distribution of residuals used to produce the forecasts (e.g., due to heteroscedasticity, non-normality, or correlation of residuals), then this is likely to make the forecast inaccurate. Assessing the residuals directly as part of the forecast accuracy assessment process is the most direct approach to diagnose this inaccuracy.

This is one example of how residuals may be analysed and presented. Alternative presentations (e.g., QQ-plots) and analyses (e.g., residuals over time, or versus individual predictors) should be considered internally to ensure the most useful model diagnostics are obtained.

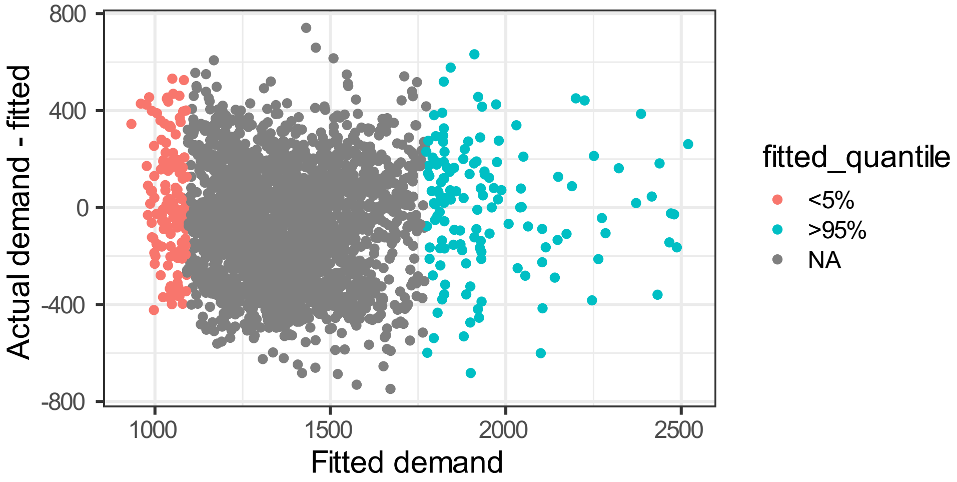


Figure 15. Residuals versus fitted values for an example forecast, with extremes of fitted values highlighted. Note: not an actual AEMO forecast.

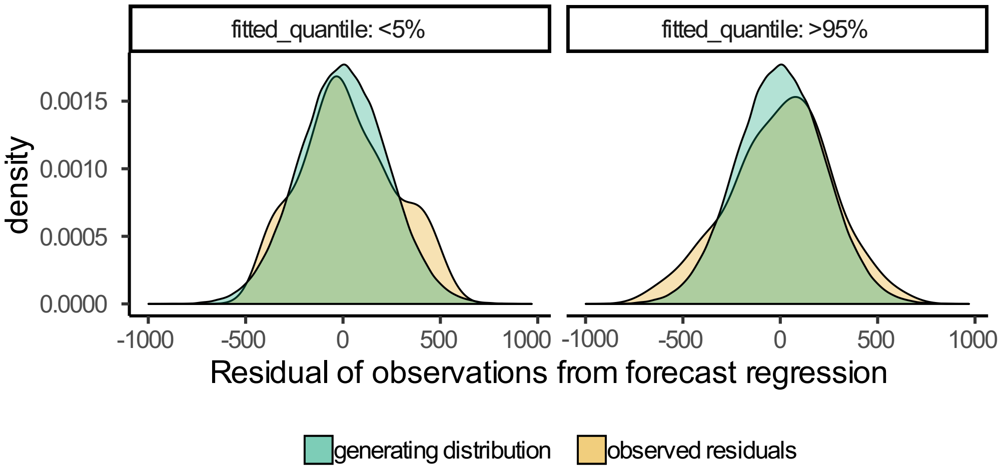


Figure 16. Comparison of kernel density estimates of the observed residuals at the fitted extremes (points highlighted in Fig. 15) versus the distribution that generated the forecast. Note that these are fitted extremes, not observed extremes; observed extremes would not have the same distribution as the generated residuals. This demonstrates that this example forecast is performing poorly. Note: not an actual AEMO forecast.

Recommendation 13. Compare the residuals that produced the simulated seasonal minimum/maximum demands, to the observed residuals from the actual minimum/maximum demand intervals. This is to assess the plausibility of forecasting the observed minimum/maximum demands.

Each observation used to generate the minimum/maximum demand forecast (from the 3,000 synthetic weather-years) had an associated forecast ($f(x)$), and a residual. Thus, the distribution of residuals at the minimum/maximum demand interval can be extracted from the forecast – for the maximum, these residuals will generally be positive, for the minimum, they will generally be negative. This distribution of residuals can then be compared to the residual of the observed minimum/maximum demand: if these do not align, it would suggest that there are problems with the residuals in the forecasting method. Fig. 17 provides an example of this comparison.

This approach, on its own, will not present a holistic picture of forecast accuracy, but provides an additional line of evidence that may assist with diagnosis of problems with minimum/maximum demand forecasts.

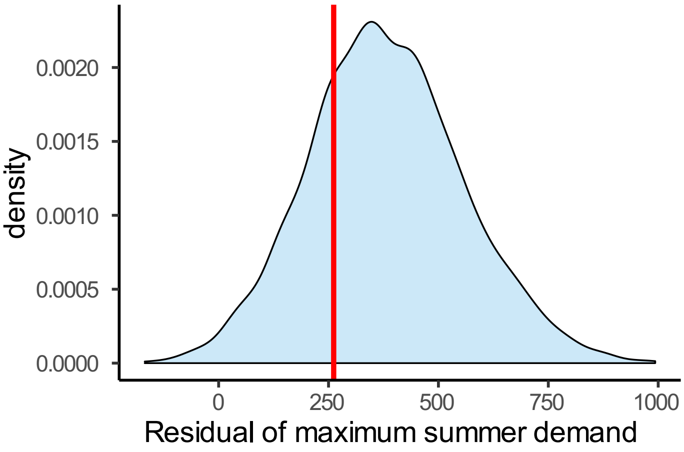


Figure 17. Example residual from forecast regression of seasonal maximum demand (red) compared against the distribution of residuals that generated maximum demand in the forecasting process. Note: not an actual AEMO forecast.

Recommendation 14. The strategies proposed in Recommendations 12 and 13 are appropriate for the existing regression-simulation forecasting framework; if or when the forecasting methodology changes, these methods should be assessed for relevance and replaced by other (model-specific) diagnostics if necessary.

Appendix A: Relative score simulations

To investigate the behaviour of the relative score function, a simple synthetic test was performed. Let the true observed value of the maximum demand be *y* = 2,500. The forecast of this maximum is a Gaussian distribution with mean *m* and standard deviation *s*, where *m* varies from 2,000 to 3,000, and *s* is either 100 or 200. Then, the relative score was used to assess (i.e., score) each of these forecasts. This process was then repeated with the modified relative score (where the relative score is normalised by the actual value *y*, rather than the forecast, q_p), and also with a scaled scenario (*y*=5,000, *m* varying from 4,000 to 6,000). The resulting relative scores appear in Fig. 18.

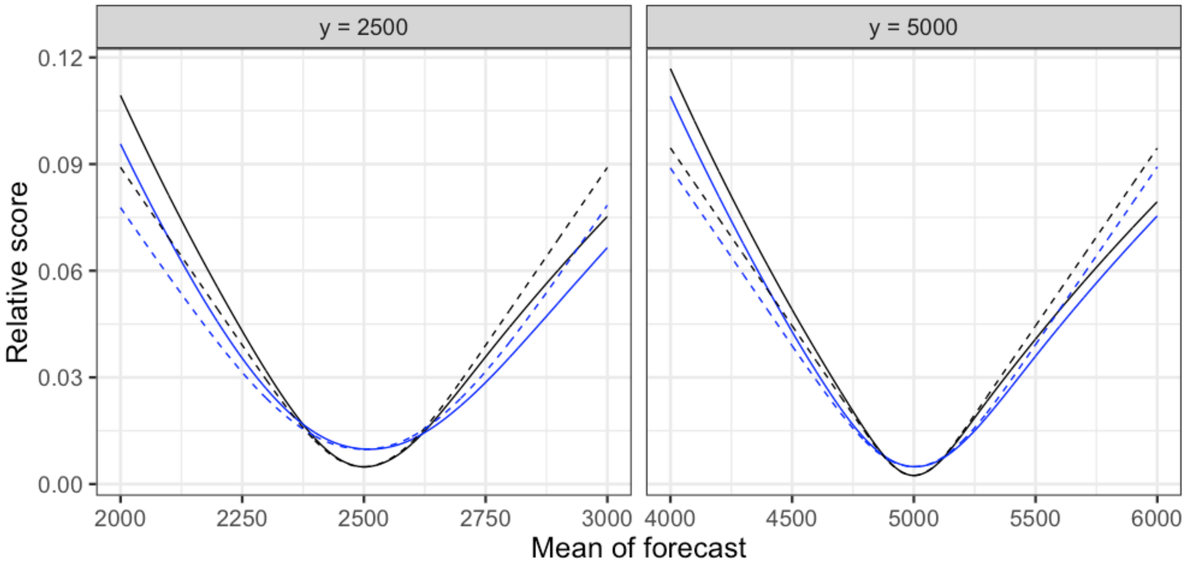


Figure 18. Example of current (solid lines) and proposed (dashed lines) relative score. The example forecast used here is a Gaussian distribution, with mean taking values along the x-axis, and standard deviation of 100 (black lines), or 200 (blue lines). This figure indicates the asymmetry of the current score formula, and demonstrates the consistency of this metric under scaling.

Appendix B: Number of observations required to detect forecast inaccuracy

To illustrate how many observations may be required to detect forecast inaccuracy, we consider a simple scenario.

Assume that we wish to count only the number of forecasts that are within the interval defined by the 90%-10% POE. By definition, 80% of observations should be within this interval. Given some number of observations, we can apply a standard frequentist statistical test, with the null hypothesis being *the forecast is correct*, that is, *80% are within the interval*. The null hypothesis is rejected if there is a probability of 5% or less of observing the data under this null hypothesis.

Now, suppose the forecast is incorrect, then we can determine the probability of detecting that error, that is, the probability that the null hypothesis is correctly rejected. Fig. 19 shows the relationship between the true probability of being within the 90%-10% POE interval and the probability of detecting an error, depending on the number of observations. For example, suppose the true probability is 0.5. With only 5 observations, the probability of detecting an error is less than 0.25. However, with 20 observations, the probability of detecting that same error is approximately 0.87. This demonstrates that to detect small errors many observations are required, but very large errors in the forecast may be detected with only 10 observations.

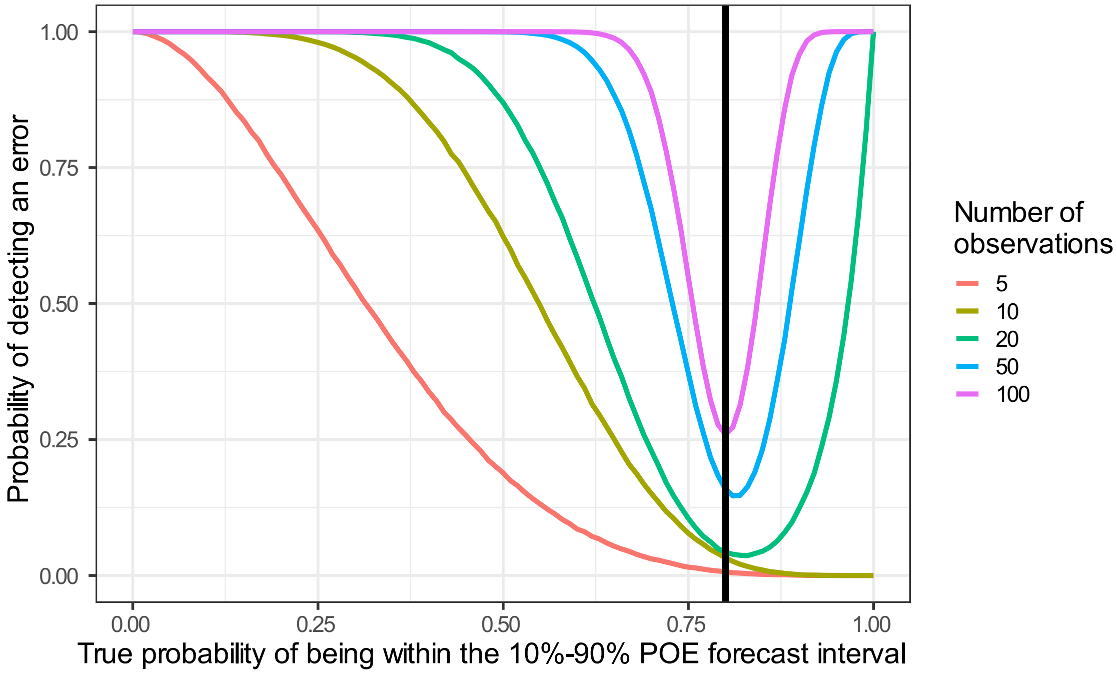


Figure 19. Probability of detecting an error in an interval forecast, based on the number of data points observed, and the magnitude of the error. In each case, the null hypothesis is that 80% of observations fall within the interval, and the test is applied at significance level 0.05.