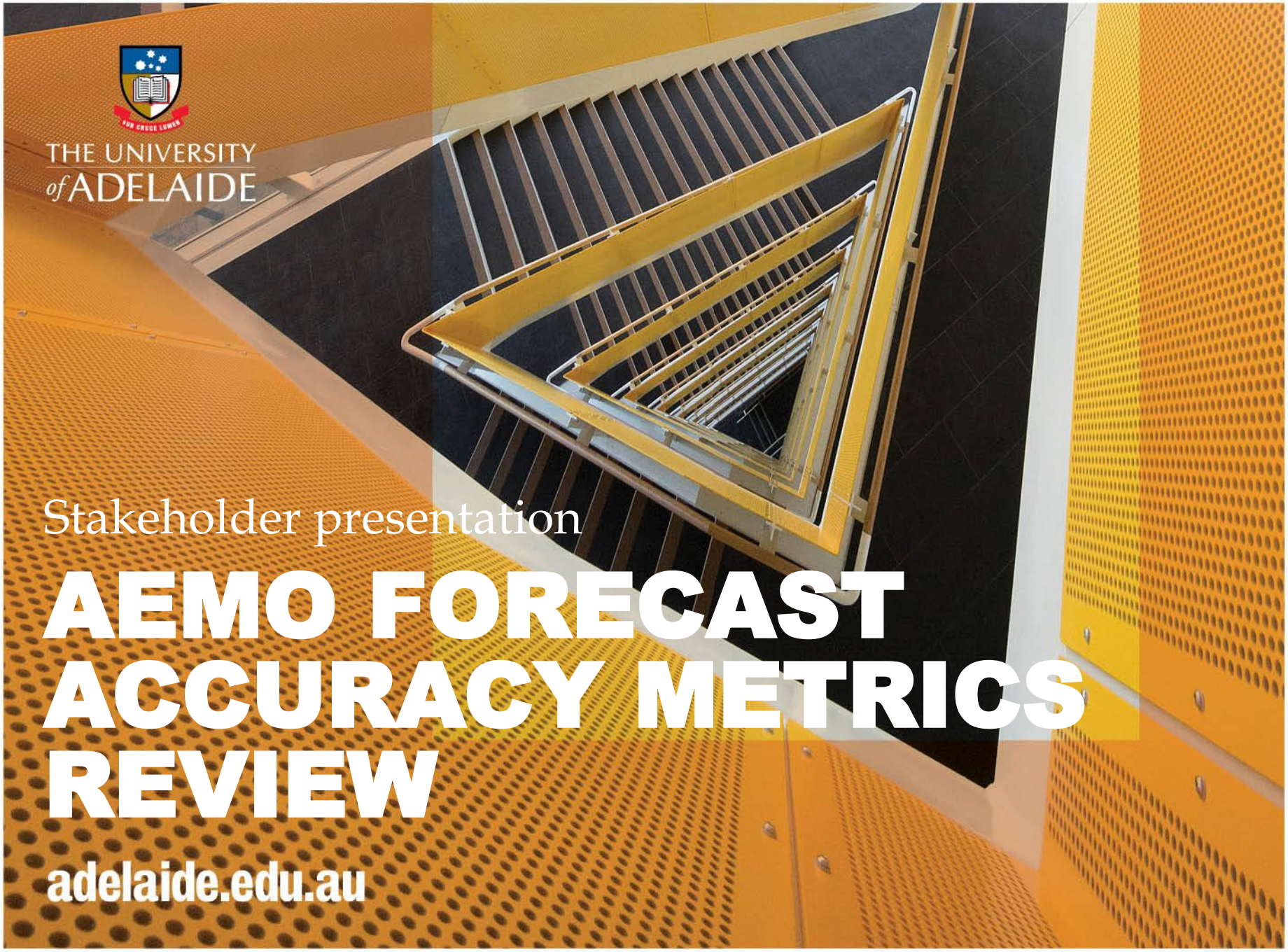


THE UNIVERSITY of ADELAIDE

Stakeholder presentation

AEMO FORECAST ACCURACY METRICS REVIEW

adelaide.edu.au

AEMO Forecast Accuracy Metrics Review

- AEMO produces forecasts of consumption and demand, and reports on the accuracy of these forecasts. The University of Adelaide team was tasked with providing an expert review of AEMO's forecast accuracy metrics.

- Broadly, current AEMO practices are appropriate and well-supported. We provide 14 recommendations for improvements to forecast accuracy reporting and monitoring.

- Recommendations 1-3 concern annual consumption, and pertain primarily to communication of results.

- Recommendations 4-14 concern peak demand, and include communication of results, review of assumptions, modifications of some approaches, and proposed new approaches.

AEMO Forecast Accuracy Metrics Review

- All recommendations are based on the 2018 FAR, and Performance Dashboard as it has been presented to us.
- All examples are for indicative purposes only.
- These recommendations are conceptual/theoretical in nature, based on the high-level information available to the University of Adelaide team; AEMO will need to assess the feasibility of producing the recommended outputs based on internal expertise regarding the forecasting process, data logistics, etc.
- Specific details around what information should be presented or prioritised is to be determined internally by AEMO and informed by stakeholders.

Annual consumption

Annual consumption

Recommendation 1. Percentage error is best practice for assessing accuracy of annual consumption. It should continue to be used for this purpose.

Recommendation 2. Wherever percentage error is reported for input quantities, include an indication of how the error in that input impacts the overall headline figure (*annual operational consumption (sent out)*).

- Example: When reporting the key input *Gross State Product* (GSP), compute the percentage error in the annual consumption using the actual value of GSP (and all other known inputs) and the forecast value of GSP (with all other known inputs).

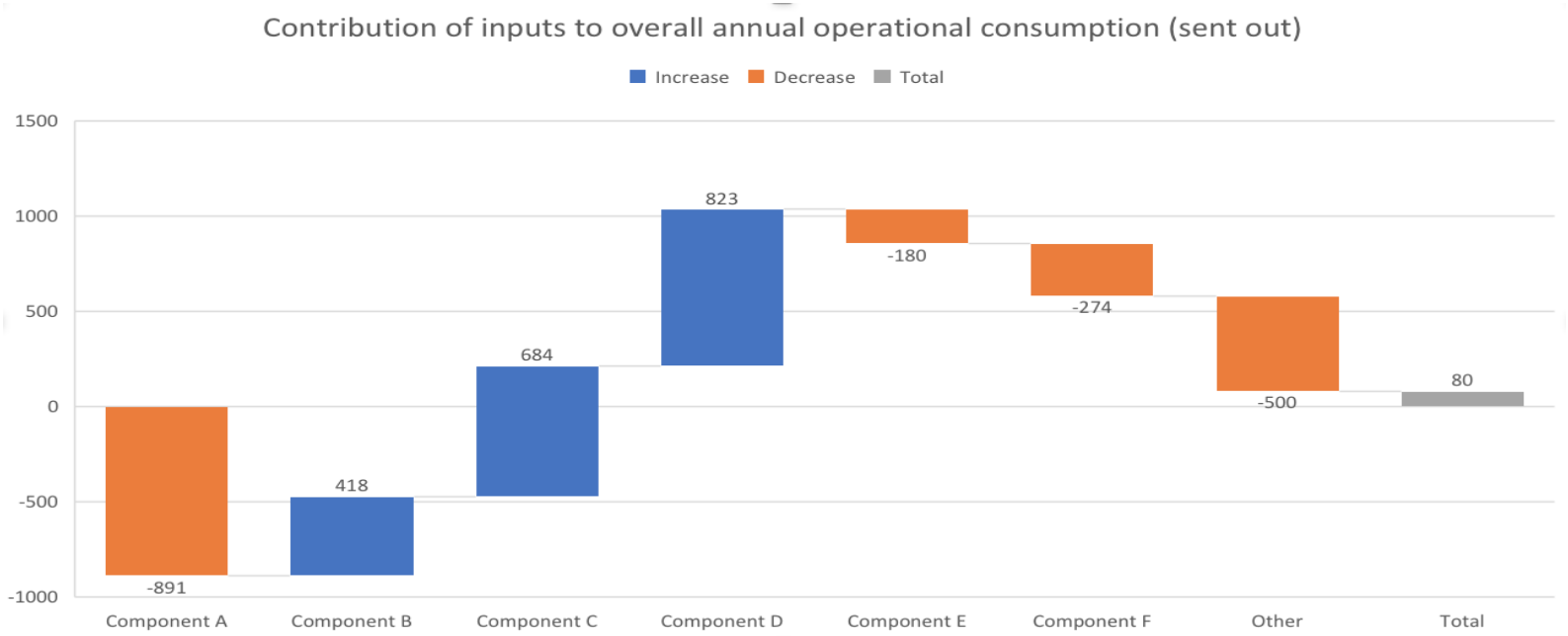
- This should be applied to all relevant inputs, including weather.

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%	x.xx%
Cooling degree days (CDD)	449	577	128	22.2%	x.xx%

Modified from Table 8, 2018 FAR.
*Note that some ''actuals'' are model estimates, that may change year on year.
Some components take months or a year to observe, for instance some inputs come from the AER submit by regulated businesses.

Annual consumption

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

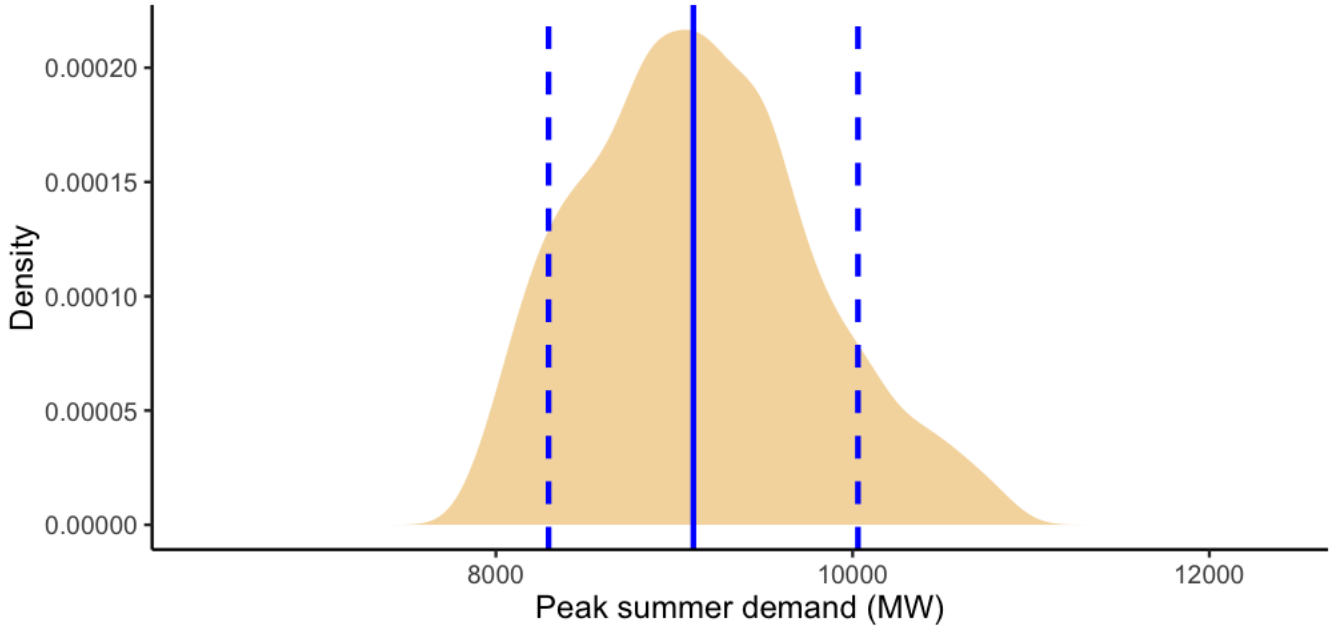


Modified from Table 8, 2018 FAR.
*Note that some ''actuals'' are model estimates, that may change year on year.
Some components take months or a year to observe, for instance some inputs come from the AER submit by regulated businesses.

Minimum/maximum probabilistic demand

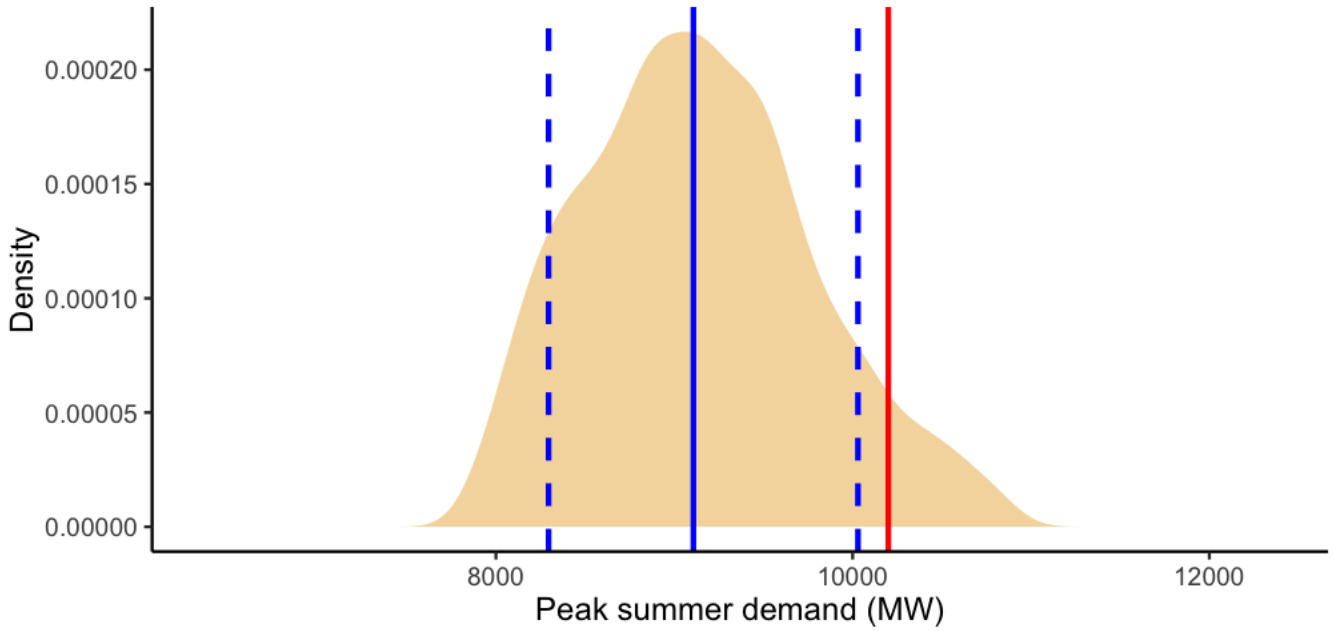
Min/max probabilistic demand - Summary

- The forecast of the peak (min/max) demand is calculated through a complex simulation process, driven by synthetic temperature-years, and including stochastic volatility. The resulting forecast is thus a probability distribution.



Assessing probabilistic demand forecasts

- As the actual peak demand in a season is a single observation, it is not possible to assess a probabilistic forecast from solely that data.
- A range of methods each provide different information.



Qualitative comparison - Review

- The primary method of reporting accuracy of forecast peak demand is *qualitative comparison*, i.e., describing where on the exceedance distribution the peak lies, and contextual factors that may explain this.

- Example: ''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 valuable context around forecast accuracy. It should continue to be used in its current form.

Probabilistic drivers

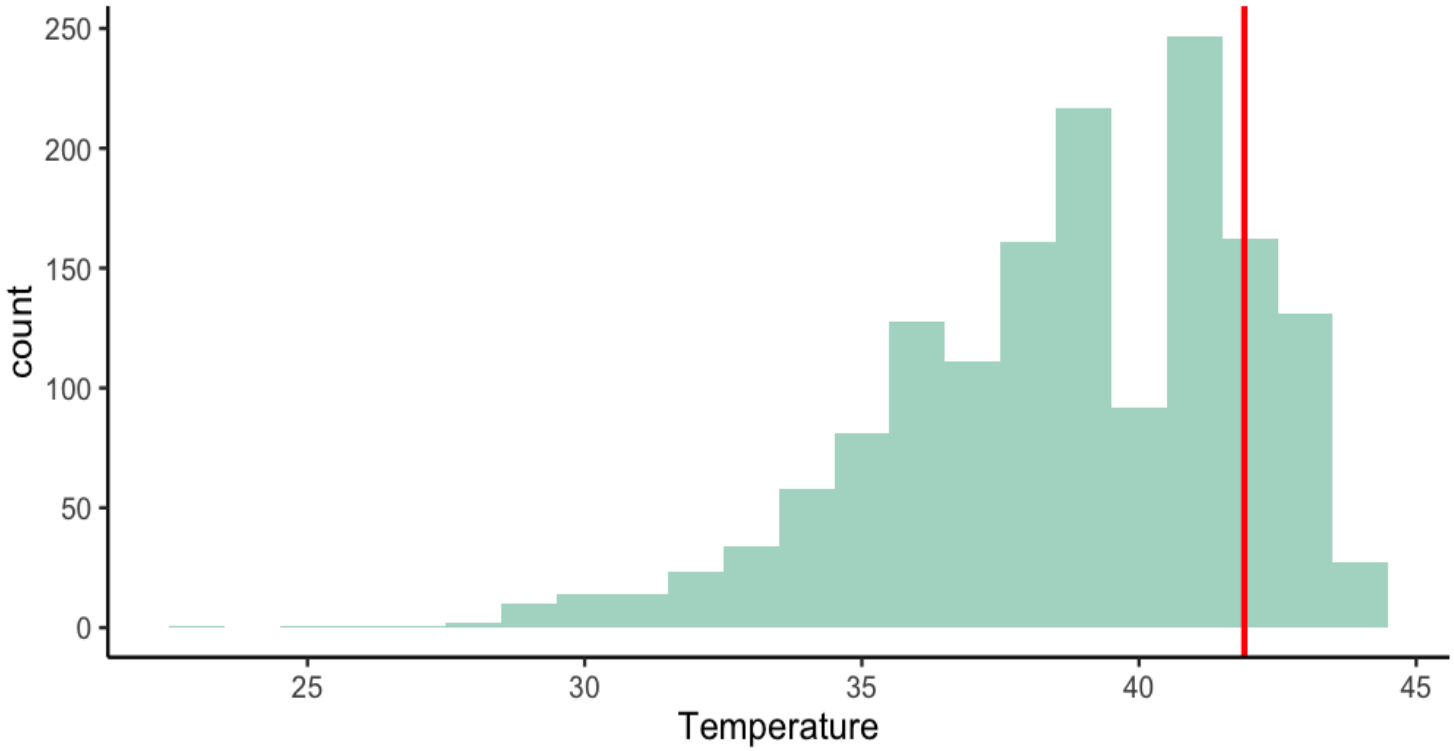
- Current approach is to present probabilistic drivers of demand in a table.

Operational maximum demand	Actual (summer)	Forecast 10% POE	Forecast 50% POE	Forecast 90% POE
Rooftop PV at time of operational max demand (MW)	369	298 to 399		
Weather – at time of operational max demand				
Temperature (°C)	31.5	37.1	32.6	31.0
Dry temperature rolling 3-day cooling degrees	7.14*			
Characteristics of peak demand day				
Time (local)	5:00 pm	3:30 pm to 4:30 pm		
Weekend	No			
Public holiday	No			
School holiday	No			

From Table 11, 2018 FAR.

Probabilistic drivers

- Recommendation 5. Present the key drivers of peak demand graphically.



Some probabilistic intuition

- Consider the simple example of a coin flip: We wish to determine if the coin is biased.
- If you flip the coin many times, it should show heads half the time (if it is unbiased).
- This is *the Law of Large Numbers*.
- If, after observing many values, that proportion of coin flips that are heads differs substantially from 50%, you may have statistical evidence that the coin is biased.
- Your ability to detect bias (with a statistical test) depends both on how many values you observe, and how bad that bias is: Something strongly biased will be much easier to detect than a small difference.
- E.g., 10 coin flips, all tails -- evidence of bias.
- 10 coin flips, 4 heads -- insufficient evidence.
- 1000 coin flips, 400 heads -- evidence of bias.

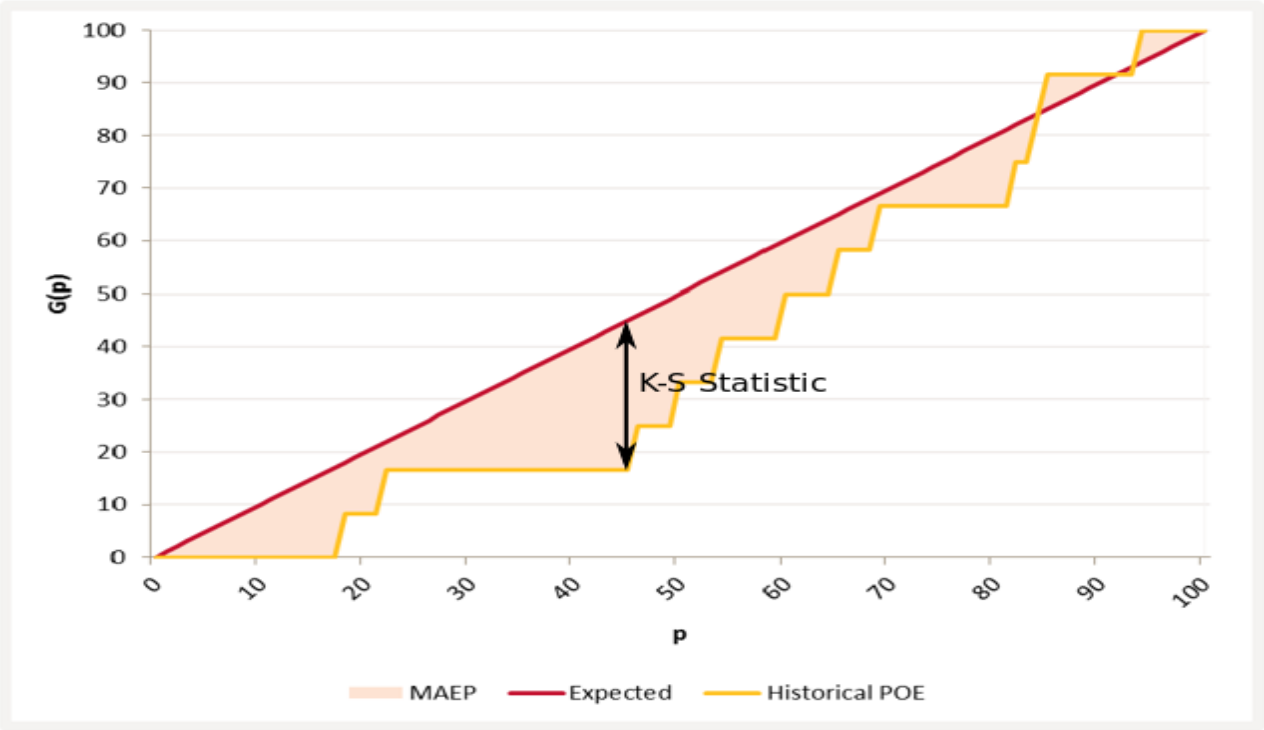
Data aggregation

- Forecast assessment is similar: Many observations may be required to detect forecast inaccuracy.
- However, those observations must be independent, and have the same appropriate distribution.
- One possible approach to producing more observations, is to disaggregate from seasonal forecasts into monthly (or even weekly) forecasts, for assessment.
- However, these disaggregated forecasts may not be independent, and likely have distributions that are different to the key forecasts that are relevant to business need.

- Recommendation 6. Assess data aggregation processes to ensure that distributional assumptions are met, in particular ensuring that aggregation occurs on a scale that is relevant to business need (i.e., seasonal max/min demand).

MAEP / Kolmogorov-Smirnov

- Given sufficiently many observations, this is a standard method for comparing distributions, and consequently is a key tool used for probabilistic forecast assessment across industries (see e.g., Gneiting, Balabdouli and Raftery 2007).



From Figure 2, 2015 FAR.

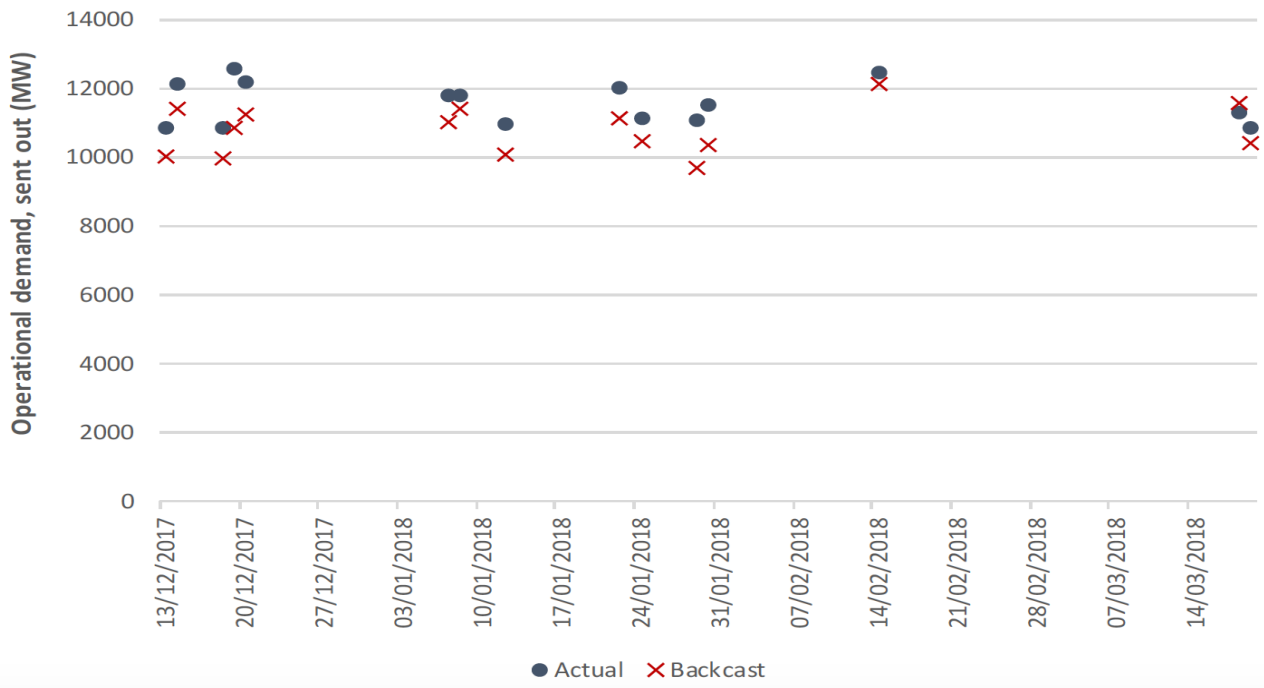
MAEP / Kolmogorov-Smirnov

- Recommendation 7. Continue to assess empirical forecast distribution fit, though metrics such as MAEP and the Kolmogorov-Smirnov statistic. Ensure that appropriate distributional assumptions are met in all analyses (as per Recommendation 6).

Backcasting

- In 2018 FAR but not informative (removed from 2019 update).
- Recommendation 8. Discontinue backcasting; replace with *full-season hindcasting* (see Recommendation 9).

Figure 12 Actual versus backcast max demand half-hour for top 15 highest demand days in New South Wales



Hindcasting / simulated history - Overview

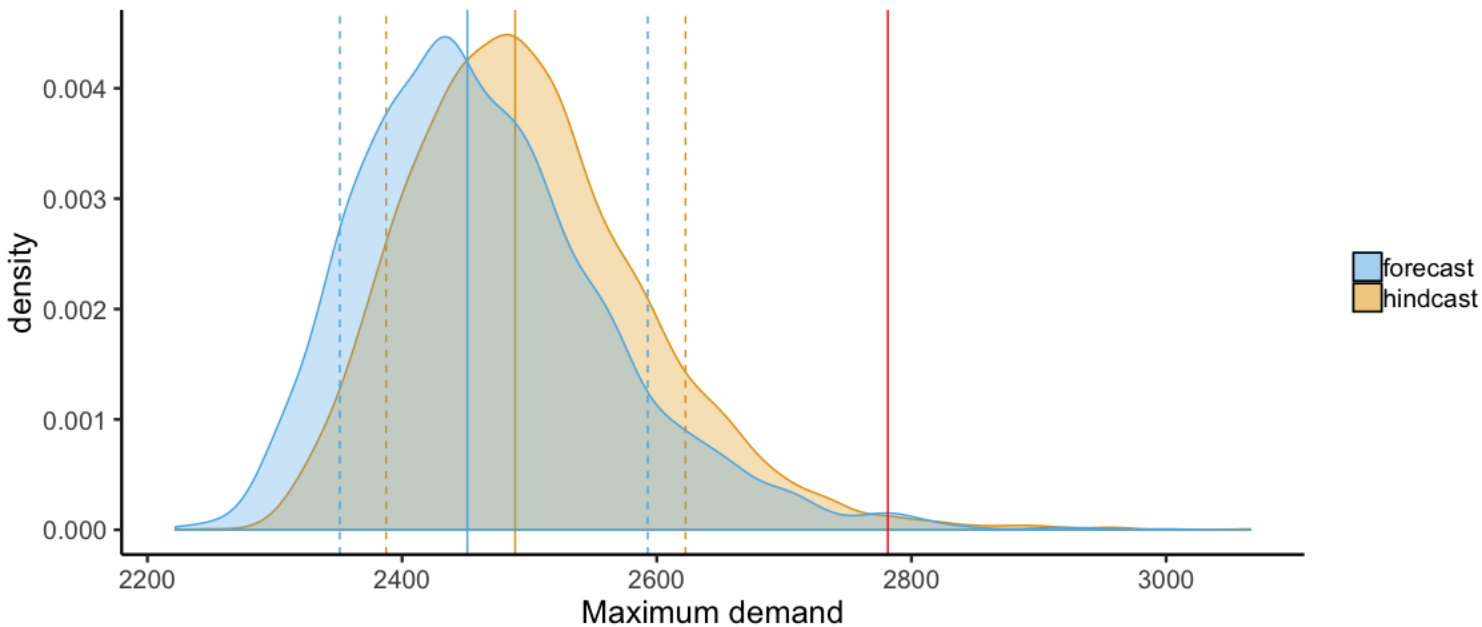
- Two approaches:

- *Full-season hindcasting*: e.g., ESOO 2018 Forecast compared against seasonal hindcast made in 2019.
  - Using actual weather from that year, actual growth drivers (e.g., PV capacity), etc.
  - Purpose: Compare the forecast that was made, to what would be made now (i.e., with actual inputs).

- *Simulated history*: Actual maximum demand, compared against hindcast applying current method to past seasons (with synthetic weather / drivers).
  - E.g., actual 2017 maximum vs. 2017 hindcast,
  - actual 2016 maximum vs. 2016 hindcast, and so on.
  - Purpose: Testing on historical data for model development.

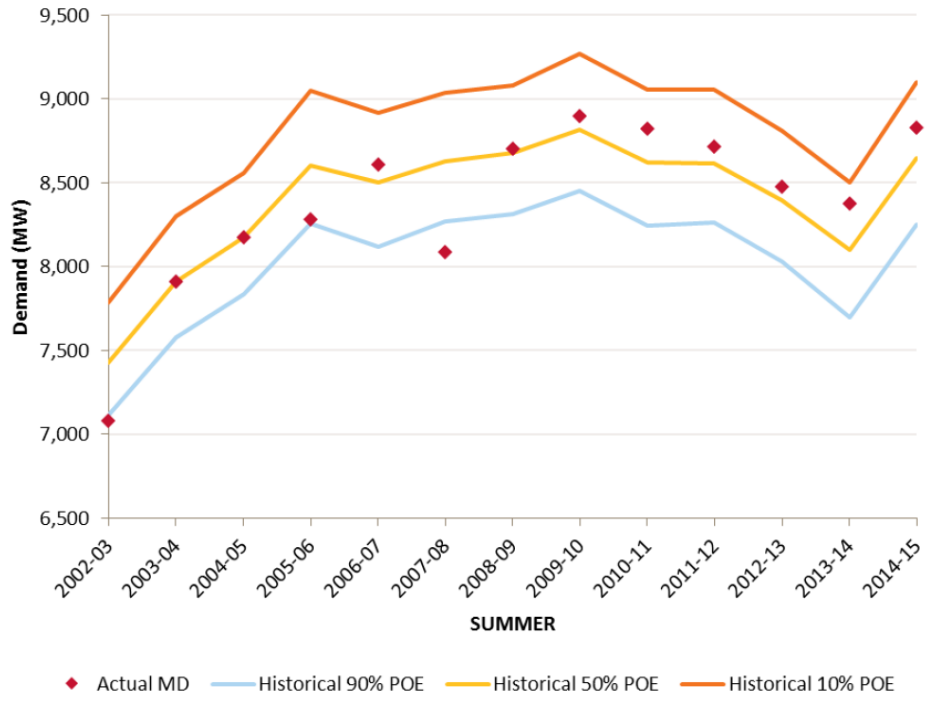
Full-season hindcasting

- Recommendation 9. Perform *full-season hindcasting* by using the actual year of temperature data in the forecasting method, in the place of the synthetic temperature-years.
- Note these should not be identical -- but differences may provide context around possible errors.



Simulated history

- This approach is to
  - apply the current forecasting model to previous years,
  - produce probabilistic forecasts of peak demand for these years, then
  - compare the actual maximum demand to those forecasts.



From Figure 6, 2015 FAR.

Simulated history

Recommendation 10.

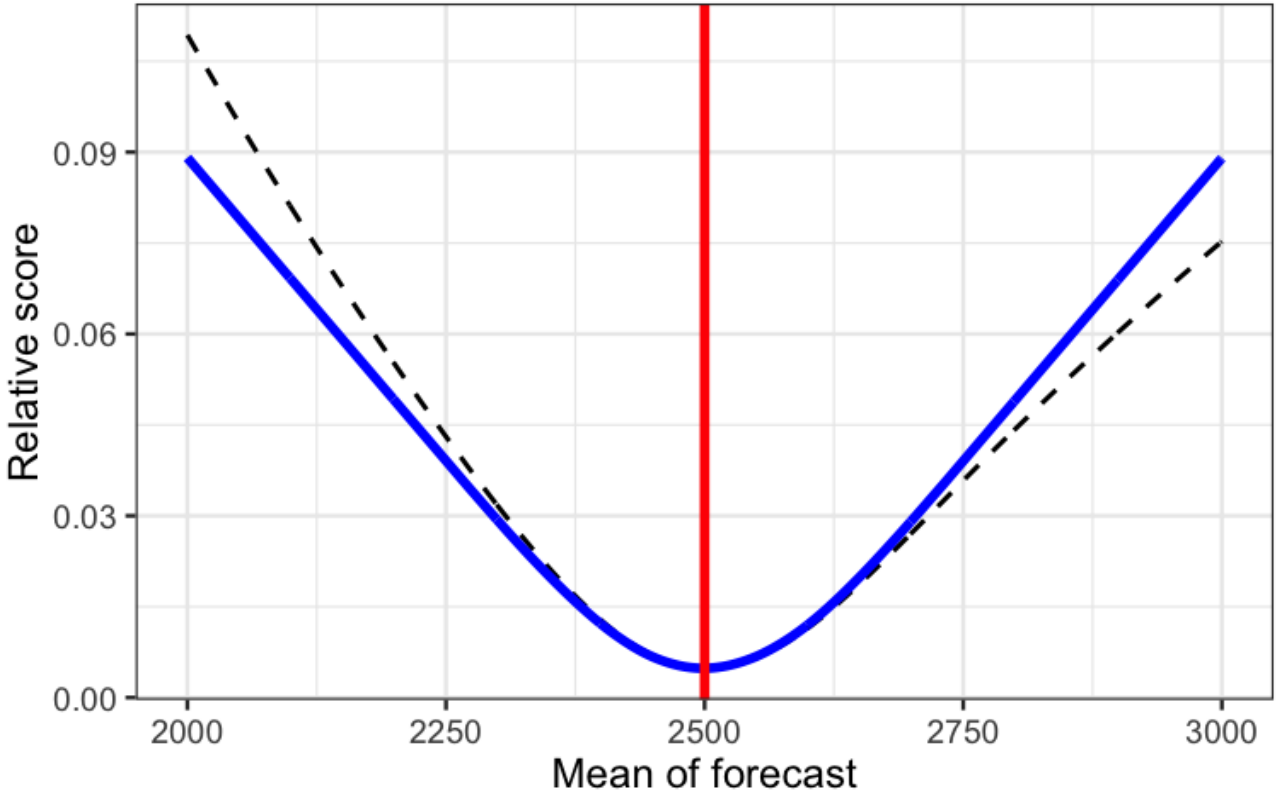
- Assess the assumption that the forecast model can be applied historically, e.g. accounting for past consumer behaviour.
- If appropriate, use these simulated historical seasonal maximum demand forecasts to compute MAEP and KS statistic, and then evaluate the appropriate statistical test for the KS statistic (if sample size is sufficient).

Relative score

- The score forecast accuracy metric calculates the average distance of the observation from the forecast quantile.
- It is used to compare forecasts, favouring those that are *accurate* and *sharp*. Smaller values indicate better performance.
- However, the score is in units relative to the forecast: This means that there is no comparison between a large region and a small region (like NSW and Tas.).
- The relative score tries to normalise the forecast accuracy metric so that they are on the same scale.
- Based on the ''pinball loss function'' – commonly used in e.g., the Global Energy Forecasting Competition (2014 & 2017).
- AEMO currently normalises the score metric by dividing by the forecast value. This approach favours overestimates.
- Recommendation 11. Modify the normalisation to remove this bias.

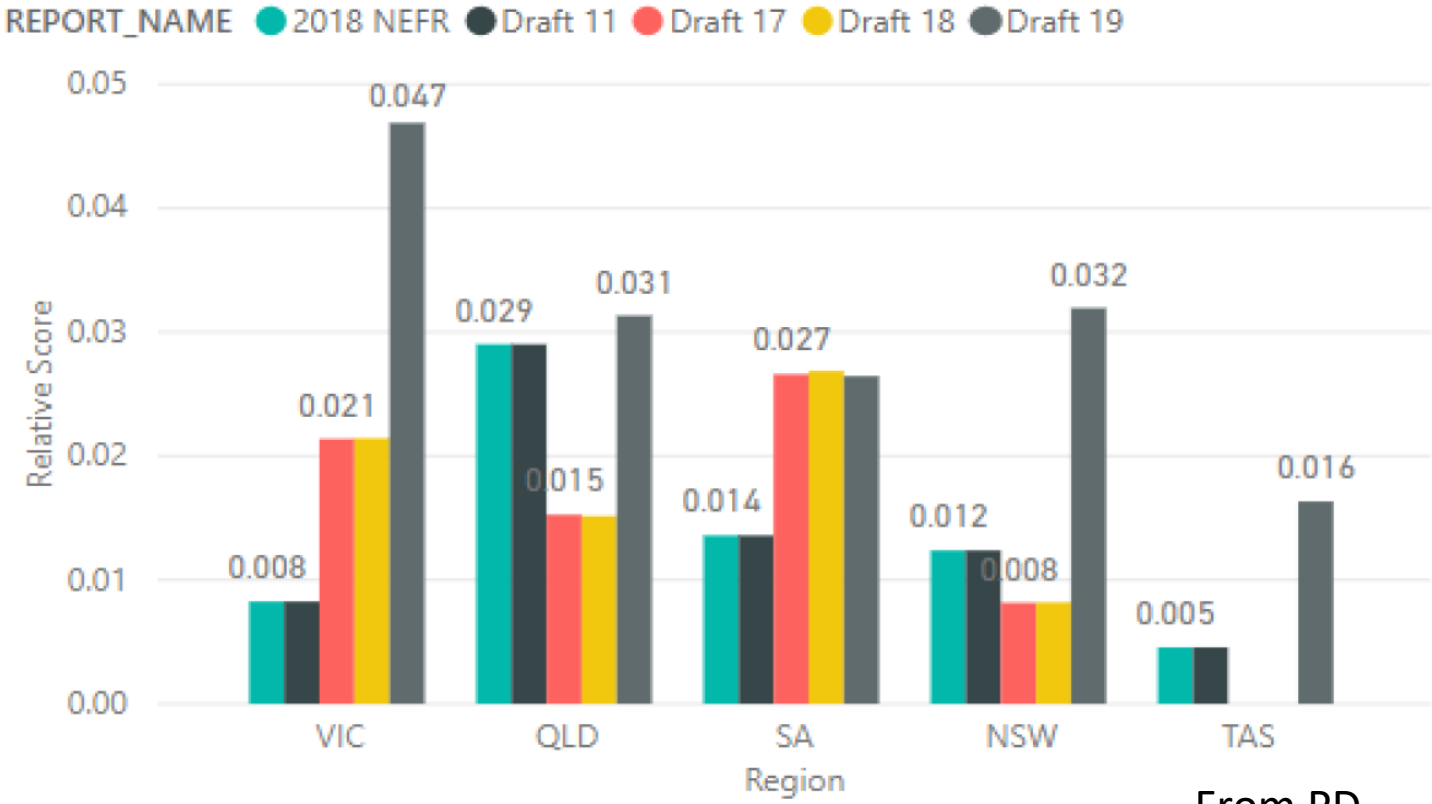
Relative score

Relative score metric divided by forecast value



Relative score

- Following Recommendation 11, this is a reasonable tool for drawing broad conclusions around forecast accuracy across datasets, and it should continue to be used for this purpose.

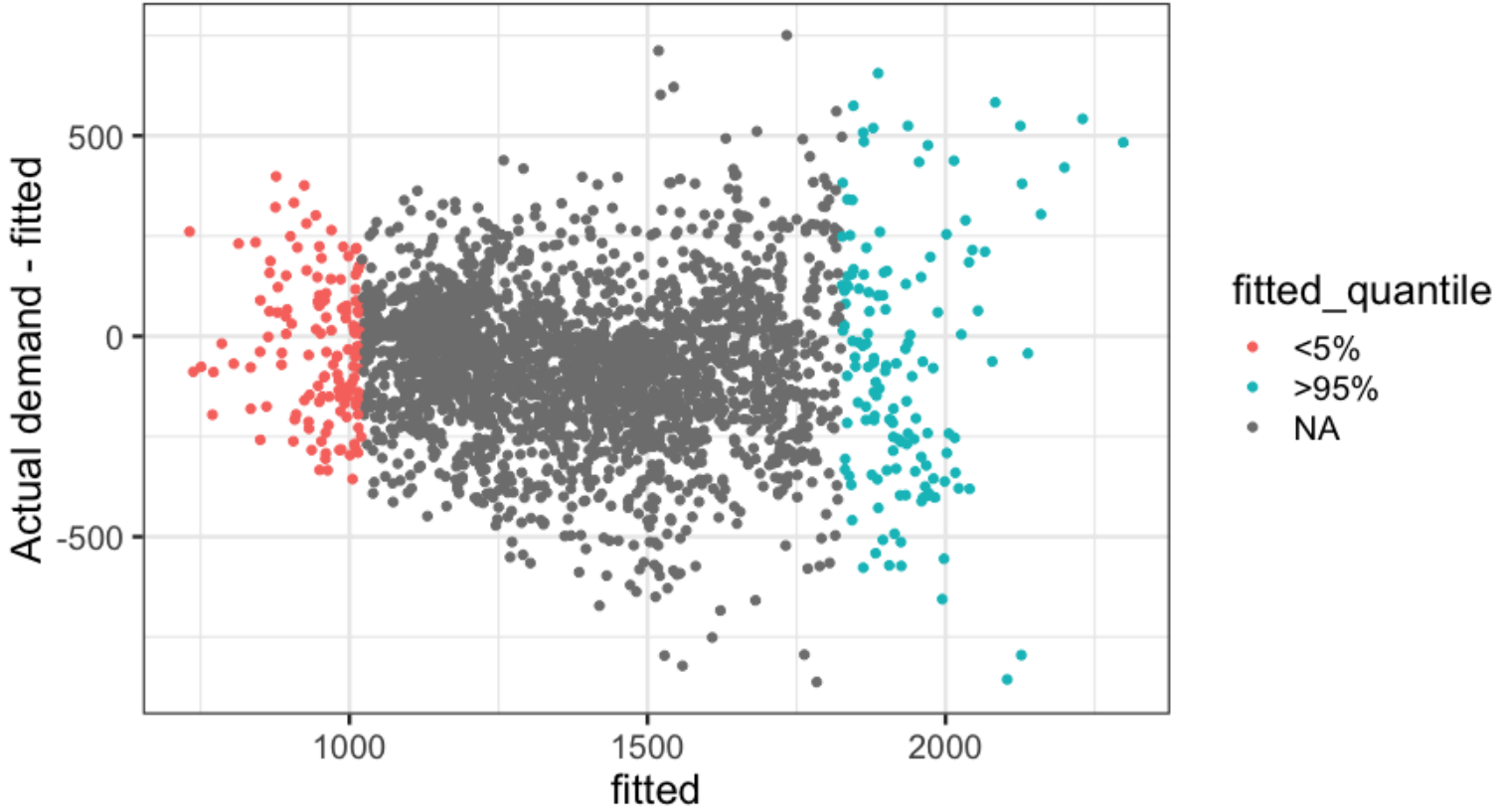


From PD.

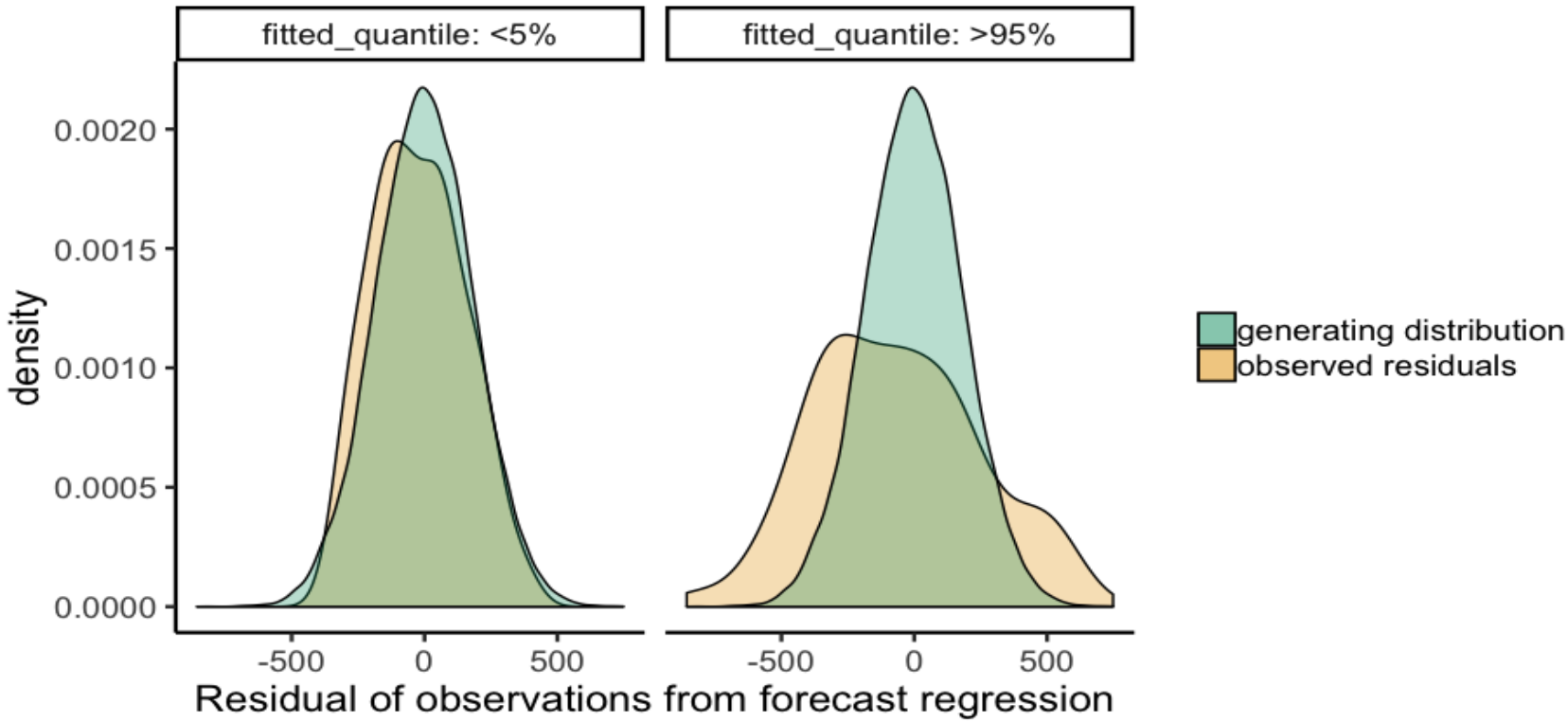
Model development accuracy and assumption metrics

Distribution of residuals

Recommendation 12. Extract the top & bottom 5% of values of the forecast regression, and analyse their observed residuals.

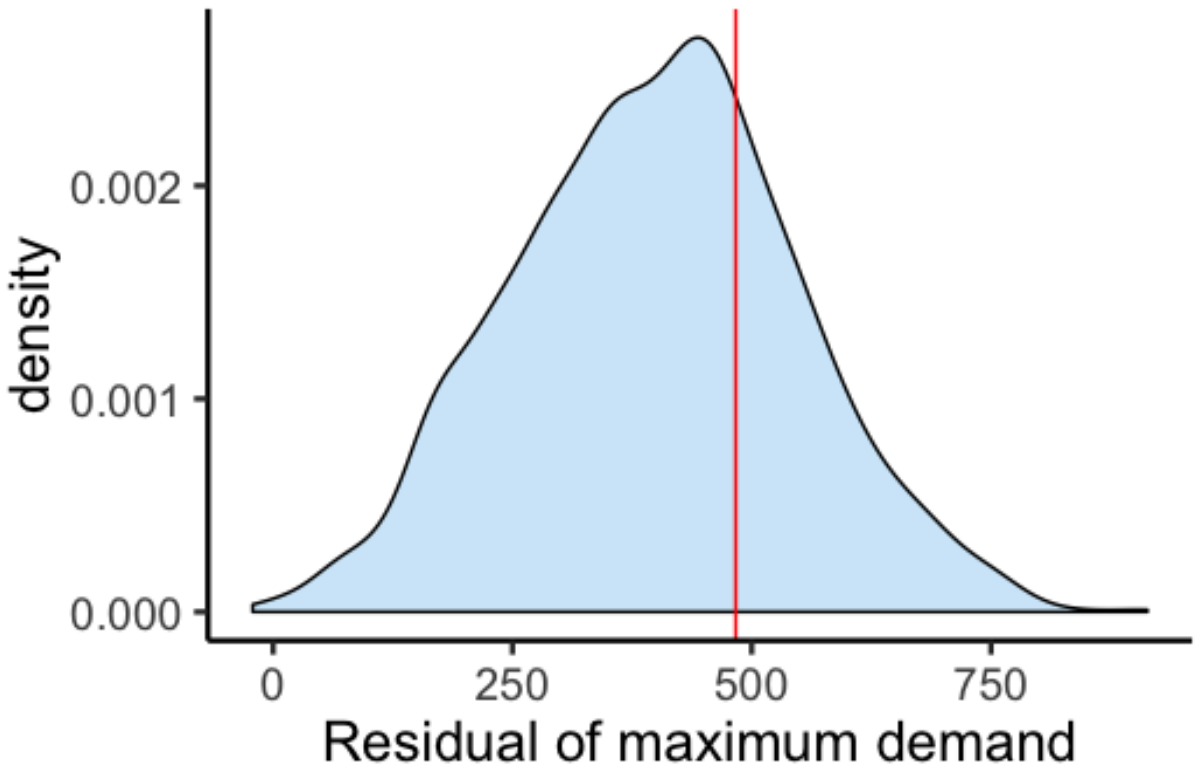


Distribution of residuals



Distribution of residuals

Recommendation 13. Extract the residuals of the peak in the forecast simulations; compare to the observed residual of the peak.



Distribution of residuals

- Recommendations 12 and 13 present two examples of model diagnostics that can provide insight when considered as part of internal forecast accuracy assessment.

- Additional model diagnostics should also be considered, particularly for inclusion in the performance dashboard, including:
  - Quantile-quantile plots,
  - Residuals versus predictors,
  - Residuals over time.

- Recommendation 14. The strategies proposed are appropriate for the existing regression-simulation forecasting framework; if/when the forecasting methodology changes, it may be that other (model-specific) diagnostics should instead be used.

Assessing min/max probabilistic demand - Summary

- Forecast Accuracy Report (2018):
  - *Qualitative comparison (R4),*
  - *Distribution of drivers (R5), and*
  - *Backcasting (R8), replaced with full-season hindcasting (R9).*
- Performance monitoring dashboard:
  - *MAEP/Kolmogorov-Smirnov statistic (R6, 7),*
  - *Simulated history (R10), and,*
  - *Relative score (R11).*
- New proposals – for performance dashboard:
  - *Analysis of residuals (R12, 13, 14).*

Next steps

- AEMO and stakeholders to provide feedback within the next week – please email any feedback to the AEMO forecasting team.

- This feedback will be incorporated into the final report, due mid-August.

SUB CRUCE LUMEN

THE UNIVERSITY of ADELAIDE

CRICOS PROVIDER NUMBER 00123M