Advice for Automation of Forecasts: A Framework

Deryn Griffiths, Harry Jack, Michael Foley, Ioanna Ioannou and Maoyuan Liu

April 2017
Advice for Automation of Forecasts: A Framework

Deryn Griffiths, Harry Jack, Michael Foley, Ioanna Ioannou and Maoyuan Liu

Bureau Research Report No. 021
April 2017
Enquiries should be addressed to:

Deryn Griffiths

Bureau of Meteorology
GPO Box 1289, Melbourne
Victoria 3001, Australia

deryn.griffiths@bom.gov.au

Copyright and Disclaimer

© Commonwealth of Australia 2017

This work is copyright. Apart from any use as permitted under the Copyright Act 1968, no part may be reproduced without prior written permission from the Bureau of Meteorology. Refer to www.bom.gov.au/other/copyright.shtml for further information.
# Contents

Abstract ... 1

1. **Introduction** ... 2

2. **Data** ... 3

   2.1 Sites and Station Groups ... 3

   2.2 Seasons ... 3

   2.3 Forecast Parameter ... 3

   2.4 Forecasts and Lead Days ... 3

   2.5 Missing Forecast Data ... 4

   2.6 Observations of Rainfall and Rain Events ... 4

3. **Methods** ... 5

   3.1 Verification Statistics Used ... 5

      3.1.1 Brier Score ... 6

      3.1.2 Relative Economic Value ... 7

      3.1.3 Accuracy ... 8

   3.2 Hypothesis Testing ... 8

4. **Results** ... 9

   4.1 Australia-wide Results for DailyPoP ... 9

   4.2 Winter Regional Results for DailyPoP ... 11

   4.3 Summer Regional Results for DailyPoP ... 15

   4.4 Users with varied sensitivities to rainfall ... 16

   4.5 Text Service ... 16

   4.6 Spring and Autumn Results for DailyPoP ... 18

   4.7 DailyPoP1 Results ... 19

   4.8 DailyPoP15 Results ... 20

5. **Further Discussion** ... 21

   5.1 Objective decision making ... 21

   5.2 Improving Confidence in Results ... 21

   5.3 Weighing risks based on impact ... 22

   5.4 Relationship to broader service ... 22

   5.5 Acceptable Skill Level ... 22

   5.6 Impacts of, and on, Future Improvements ... 22

6. **Conclusion** ... 24

Acknowledgements ... 25

References ... 25

Appendix 1 Station groups ... 26

Appendix 2 Sub-region Scorecards ... 30
List of Tables

Table 3.1  Contingency table for a binary forecast. ................................................................. 7
Table 4.1  Scenarios of skill that may be lost by automating DailyPoP forecasts. 
                      Australia-wide results. ................................................................. 10
Table 4.2  Scenarios of skill that may be lost by automating DailyPoP forecasts in winter. 
                      Regional results. ........................................................................... 11
Table 4.3  Scenarios of skill that may be lost by automating DailyPoP forecasts in summer. 
                      Regional results. ........................................................................... 15
Table 4.4  Scenarios of skill that may be lost by automating DailyPoP15 in winter. 
                      Regional results. ........................................................................... 20
Table 4.5  Scenarios of skill that may be lost by automating DailyPoP15 in summer. 
                      Regional results. ........................................................................... 20
Table A.1  Scenarios of skill that may be lost by automating DailyPoP forecasts in winter. 
                      Sub-group results. ........................................................................ 30
Table A.2  Scenarios of skill that may be lost by automating DailyPoP forecasts in summer. 
                      Sub-group results................................................................. 31

List of Figures

Figure 4.1  DailyPoP Brier Score and Difference in Brier Score by lead day. 
                      Australia-wide results. ........................................................................ 9
Figure 4.2  DailyPoP Brier Score and Difference in Brier Score by lead day for winter 2016. 
                      Regional results. ........................................................................ 13
Figure 4.3  DailyPoP Brier Score and Difference in Brier Score by lead day for summer 2015-16. 
                      Regional results. ........................................................................ 14
Figure 4.4  Relative Economic Value as a function of cost-loss ratio for DailyPoP lead day +5. 
                      ........................................................................ 16
Figure 4.5  Icon Accuracy by lead day for summer 2015-16. Australia-wide results. 
                      ........................................................................ 17
Figure 4.6  Relative Economic Value contrasting binary and probabilistic forecasts. 
                      ........................................................................ 18
Figure 4.7  Brier Score and Difference in Brier Score by lead day for winter 2016. 
                      Greater Sydney results. ........................................................................ 19
Figure A.1  Stations used, shown by region. ........................................................................ 26
Figure A.2  Stations used in Tasmania, shown by sub-group. 
                      ........................................................................ 26
Figure A.3  Stations used in Victoria, shown by sub-group. 
                      ........................................................................ 27
Figure A.4  Stations used in NSW & ACT, shown by sub-group. 
                      ........................................................................ 27
Figure A.5  Stations used in South Australia, shown by sub-group. 
                      ........................................................................ 28
Figure A.6  Stations used in Western Australia, shown by sub-group. 
                      ........................................................................ 28
Figure A.7  Stations used in Northern Territory, shown by sub-group. 
                      ........................................................................ 29
Figure A.8  Stations used in Queensland, shown by sub-group. 
                      ........................................................................ 29

iv
ABSTRACT

This report provides an objective framework to support decisions around the automation of existing manual forecast processes. The framework is demonstrated by example, using forecasts of Chance of Rain.

We analyse the difference in quality between automated and manual forecasts of Chance of Rain at observation points in Australia for two seasons. We use the Brier Score as the primary measure of quality, and report the difference in terms of days of skill that would be lost by adopting the automated guidance. We provide confidence intervals to indicate the uncertainty in the verification.

The results show that automating winter forecasts is likely to provide an improvement in quality as measured by the Brier Score at lead days of +5 to +7 with only minimal risk of deterioration in quality. However, automating summer forecasts, especially in the tropics in the wet season, is likely to cause deterioration in the Brier Score. To assist decision makers, we quantify the possible deterioration arising as a result of automation using “likely” and “worst case” scenarios.

We discuss other aspects, some subjective, which might influence a decision regarding automation of forecasts. In particular we highlight icon and text forecasts, which are important and considered by manual forecasters, but less amenable to objective assessment than the numerical forecasts.
1. INTRODUCTION

The principal aim of this report is to provide an evidence based framework to support decision making regarding the automation of parts of the Bureau of Meteorology’s forecast services. We illustrate the framework by applying it to forecasts of Chance of Rain.

Our work is reported in the context of improving efficiency. That is, the organisation expects to automate forecasts as long as there is good confidence that doing so will not degrade the quality of forecasts as assessed over an appropriate period with appropriate measures. Indeed, the organisation may be prepared to automate forecasts even if doing so causes a level of degradation in quality as long as it deems the degradation acceptable. A reason for such a decision would be the extra value forecasters could add to other parts of the service with the time freed by automation.

To illustrate the framework we provide advice regarding automating \textit{DailyPoP} forecasts. \textit{DailyPoP} refers to the chance of receiving at least 0.2mm of rainfall at an observation point in a 24 hour period. The \textit{DailyPoP} forecast is particularly prominent in the public service provided by the Bureau of Meteorology where it is referred to as the ‘Chance of any rain’. We include brief comments on forecasts of \textit{DailyPoP1} and \textit{DailyPoP15}, the chance of receiving at least 1 mm and 15 mm of rainfall respectively.

Our general approach is to compare the past performance of automated guidance with the manually produced forecasts. We only assess the numerical quality of the forecasts. This is a limitation. We do not assess the quality of the text or other downstream products. There is anecdotal evidence from forecasters that they modify the manual forecasts to produce desired downstream products. Exploring the extent of such behaviour, and identifying tensions in the service definition that encourage such behaviour is beyond the scope of this report. Indeed, attribution for the reasons of the relative performance of the automated and manual forecasts is outside the scope of this report. We focus on \textit{what} the difference is, rather than \textit{why} there is a difference.

A further limitation is our assumption that comparing past performance is a good estimate of future performance. We do not allow for future improvements to automated or manual forecast processes or the impact of annual climate variability on performance.

We report ‘likely’ and ‘worst case’ scenarios in terms of days of skill that may be lost by using the automated forecast. If the ‘worst case’ is that we gain skill by automating, we assume that automation is acceptable. In doing so we make an assumption about the suitability of the 97.5 per cent confidence level used to create the ‘worst case’ scenario. On the other hand, we make no assumptions about what might be an acceptable level of degradation in quality, if any.

The results could be calculated to suit any supplied confidence level. Alternatively, if an ‘acceptable level of degradation’ was supplied, we could report on the confidence we have of the automated forecast performing to within that standard.

Future work is expected to extend to other parameters and may be tailored to policy decisions regarding a preferred confidence level or an ‘acceptable degradation’ threshold. Furthermore, as the automated guidance improves, the results contained within this report should be reviewed.
Section 2 discusses the data used in the analysis. Section 3 elaborates on the verification techniques. Results are presented Section 0 and Appendix 2. Discussion of the results is contained in Section 5 and summarised in Section 6.

2. DATA

This section describes the forecasts and observations used in our assessment.

2.1 Sites and Station Groups

Forecast and observation data were collated for 434 automatic weather stations around Australia.

The verification statistics were run over sub-groups of these stations. In this report we include:

- An Australia-wide group containing all 434 stations
- A sub-group for each region containing all the stations in that region. The regions are based on Australian states and territories as shown in Appendix 1.
- A set of sub-groups within each region with approximately 10 to 20 stations in each group as detailed in Appendix 1.

2.2 Seasons

We analysed data in three-month blocks.

Winter 2016 refers to June, July and August 2016, the middle of the dry season in the tropics.

Summer 2015-16 refers to December 2015 and January and February 2016, the middle of the wet season in the tropics.

2.3 Forecast Parameter

DailyPoP is the chance of rainfall, hail or snow of at least 0.2 mm (melted equivalent for hail and snow) at a rain gauge within a specified 24-hour period.

DailyPoP1 is the chance of rainfall, hail or snow of at least 1 mm (melted equivalent for hail and snow) at a rain gauge within a specified 24-hour period. Similarly, DailyPoP15 is the chance of rainfall, hail or snow of at least 15 mm.

2.4 Forecasts and Lead Days

Forecasters work within software called the Graphical Forecast Editor (GFE). Within the GFE forecasters have access to the previously issued forecast, direct numerical weather prediction model output, and the OCF consensus guidance (described below). They have tools to help them combine
ADVICE FOR AUTOMATION OF FORECASTS: A FRAMEWORK

and manipulate this information. Two types of forecasts were extracted from GFE gridcells containing the stations of interest.

A lead day +1 forecast refers to the 24-hour period starting at 15 UTC immediately after the forecast is available or issued.

**Official** refers to the afternoon issue of the forecasts as published by Bureau of Meteorology Regional Forecasting Centres. These forecasts are the manual forecasts considered in this report.

Official forecasts are used to generate many text forecasts, which can be edited by forecasters prior to being issued by the Bureau of Meteorology. Official forecasts are ingested into the Australian Digital Forecast Database and made available publically via ftp as detailed in the Product Catalogue for Real-time Data Services available at [http://www.bom.gov.au/other/charges.shtml](http://www.bom.gov.au/other/charges.shtml). Official forecasts are typically issued around 05 to 09 UTC depending on the time zone. They are issued for the following seven days.

**OCF** refers to the 18 UTC issue of the Gridded Operational Consensus Forecast (OCF). OCF is the automated guidance considered in this report. The choice of 18 UTC base time was made to ensure we were considering a forecast that would be available reliably if used for an afternoon issue of a forecast. Typically the 18 UTC guidance is available around 00 UTC.


OCF is calculated on a 0.5° by 0.5° latitude-longitude grid (roughly 50km by 50km). This grid is remapped and interpolated when ingested into the GFE. The values of OCF used for this report are those that appeared to forecasters in the GFE, interpreted as a forecast for point locations.

The OCF algorithm is under continuous development and has been upgraded since this data was collected (BOM, 2016).

### 2.5 Missing Forecast Data

If OCF data were missing, then the corresponding data in Official were discarded and vice-versa. This method ensured that Official and OCF were being compared on exactly the same data set. A combination of this and missing observations meant that 95 per cent of forecasts were available for analysis. Missing forecast data was usually due to limitations of the archival process and as such we do not expect it to have introduced a systematic bias to the data considered.

### 2.6 Observations of Rainfall and Rain Events

One minute tipping bucket rain gauge data from automatic weather stations was extracted from the Australian Data Archive for Meteorology (ADAM) and the accumulated rainfall over the 24-hour
period of the forecast calculated. Data were discarded for daily periods missing more than 30 minutes of data. Using this technique 99 per cent of daily observations was available for analysis.

In assessing DailyPoP, observations of 0.2 mm (the smallest recorded increment) or more during the forecast period were treated as a ‘wet’ event, and 0 mm as ‘dry’. Observations treated in this way are referred to as unfiltered.

In assessing DailyPoP we computed the verification statistics on filtered and unfiltered data. The filtered data was generated by discarding days with observations of exactly 0.2mm over the forecast period. That is, observations of exactly 0.2mm were treated as if they were missing.

The reason for considering filtered data for the DailyPoP is concern that observations of exactly 0.2 mm may not reflect a day with rainfall. The report may have been from dew, frost or fog which is considered precipitation (WMO 2014) but is not part of the DailyPoP forecast. Reporting on filtered and unfiltered data provides some information about the sensitivity of our results to these reports of exactly 0.2 mm.

Similarly, in assessing DailyPoP1, observations of 1 mm or more during the forecast period were treated as an event, and less than 1 mm as the event not having occurred.

The same technique was used in assessing DailyPoP15. Observations of 15 mm or more during the forecast period were treated as an event, and less than 15 mm as the event not having occurred.

3. METHODS

The verification statistics calculated are stratified by station group and by season. We group forecasts by lead day. That is, we do not mix forecasts for tomorrow with those for a week in advance when calculating the verification statistics.

3.1 Verification Statistics Used

The Brier Score is the primary metric used to evaluate the Official and OCF forecasts. It is widely used as a measure of accuracy for probabilistic forecasts due to several desirable properties (Wilks, 2011). However, a single summary metric cannot capture all aspects that may be of interest. This paper also refers to Relative Economic Value and Icon Accuracy to illustrate aspects of the forecasts which may assist in informing the organisation of risks and benefits of any decision to automate DailyPoP forecasts.

Each verification statistic for DailyPoP is calculated based on the filtered and unfiltered data sets due to uncertainty of the truth as explained at Section 2.6. It is likely that the true results for DailyPoP forecasts lie between those calculated based on the filtered and unfiltered data.
3.1.1 Brier Score

The Brier Score takes a large number, $N$, of forecast/observation pairs, $(f_t, o_t)$, and calculates the mean squared error to produce a single number between 0 (perfect forecast) and 1 (worst possible forecast).

$$
\text{Brier Score} = BS = \frac{1}{N} \sum_{t=1}^{N} (f_t - o_t)^2
$$

The Brier Score is proper (unable to be hedged) and rewards reliability (good calibration) and resolution (ability of the forecasts to discern periods with distinct relative frequencies).

To compare OCF and Official for a particular lead day we compute the Difference in Brier Score, considering all stations and an entire season,

$$
\Delta BS = BS_{OCF} - BS_{Official}
$$

Brier Score Confidence Intervals

A 95 per cent confidence interval is calculated around $\Delta BS$ using the following technique.

Given a lead-time and a particular station group, we calculate the difference in Brier Score, $\Delta BS$, for each day. Within a season we treat these daily values of $\Delta BS$ as independent samples from a population with a t-distribution. We use Student’s t-test to construct confidence intervals for $\mu$, the population mean value of the difference in the Brier Score, and to determine the statistical significance of the observed difference, $\Delta BS$.

Grouping the data from various stations by day is a conservative approach to adjust for inter-site dependence (Candille, 2007). Where the analysis included fewer than 10 rain-days or the Brier Score did not improve with lead day we have excluded the results as unlikely to be statistically meaningful.

The choice of a 95 per cent confidence interval was arbitrary. See Section 5.1 for discussion in the context of decision making.

Brier Score Improvement by Lead-Day

We fit a line to the Official Brier Score calculated for each season and station group using least-squares regression. The gradient, $\lambda$, estimates the improvement in Brier Score per lead day as the forecasts go from day + 7 to day + 1. The improvement per lead day is calculated separately for filtered and unfiltered results.

Construction of Middle and “Worst Case” Scenarios

The improvement per lead day provides context for the magnitude of the difference in Brier Score. It also allows us to convert a difference in Brier Scores, $\Delta BS$, to $\Delta BS/\lambda$, a difference in days of skill we may gain or lose by automating using OCF.
The middle scenario provided in this report is the average of the two values $\Delta BS/\lambda$ calculated based on filtered and unfiltered data.

$$\text{Middle Scenario} = \frac{1}{2} \left( \frac{\Delta BS_{\text{filtered}}}{\lambda_{\text{filtered}}} + \frac{\Delta BS_{\text{unfiltered}}}{\lambda_{\text{unfiltered}}} \right)$$

We take the upper bound of each confidence estimate, $\Delta BS + \epsilon$, and express it in terms of days of skill, $(\Delta BS + \epsilon)/\lambda$. We report the maximum of the filtered and unfiltered values as the “worst case” scenario. It shows the lead days of skill that may (in an extreme scenario) be lost by automating using OCF.

$$\text{"Worst Case" Scenario} = \max \left( \frac{(\Delta BS + \epsilon)_{\text{filtered}}}{\lambda_{\text{filtered}}}, \frac{(\Delta BS + \epsilon)_{\text{unfiltered}}}{\lambda_{\text{unfiltered}}} \right)$$

### 3.1.2 Relative Economic Value

If a perfect forecast would save a user $x$ over a climatological forecast, then the relative economic value (REV) of a real-world forecast is the fraction of $x$ that is realisable by using that forecast. It is based on a simple decision-making model with a decision being made at a set lead time. We provide some detail here for interested readers with further detail available at Richardson (2000).

The REV of a forecast depends on the cost/loss ratio of the user, where the cost is the expense of taking preventative action against inclement weather, and the loss is the saving experienced by taking preventative action when the inclement weather occurs.

To calculate the REV for a cost/loss ratio of $\alpha$ we convert probabilistic forecasts to binary forecasts, treating forecasts of at least $\alpha$ as “wet” forecasts and forecasts below $\alpha$ as “dry” forecasts. Then we obtain a contingency table for the binary forecast as per Table 3.1.

<table>
<thead>
<tr>
<th>Observation</th>
<th>Wet</th>
<th>Dry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forecast</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wet</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wet</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1 Contingency table for a binary forecast. Each value, $a$ to $d$, indicates the number of forecast-observation pairs in that category.

If a user takes preventative action when the probability exceeds their cost/loss ratio, the optimum strategy when using a perfectly reliable forecast, the Relative Economic Value is calculated as

$$\text{REV} = \frac{\min\{a, \bar{\alpha}\} - Fa(1 - \bar{\alpha}) + H\bar{\alpha}(1 - a) - \bar{\alpha}}{\min\{a, \bar{\alpha}\} - \bar{\alpha}a}$$

where, $F = c/(a + c)$, the false alarm rate or Probability of False Detection, and $H = d/(b + d)$, the hit rate or Probability of Detection and $\bar{\alpha} = (b + d)/(a + b + c + d)$, the observed frequency of the event.
3.1.3 Accuracy
For a contingency table as at Table 3.1, the Accuracy, reported as a proportion or a percentage, is calculated as

\[
\text{Accuracy} = \frac{(a + d)}{(a + b + c + d)}
\]

3.2 Hypothesis Testing
Confidence intervals, as described in Section 3.1.1, and Hypothesis Testing, are strongly related (Wilks 2011). Both provide information about \( \mu \), the population mean value of the difference in Brier Scores, \( \Delta BS = BS_{OCF} - BS_{Official} \), to inform a decision.

In the language of hypothesis testing we could choose a null hypothesis of the form

\( H_0: \mu \geq \delta \) (Official is 'significantly' better than OCF)

and an alternative hypothesis

\( H_1: \mu < \delta \) (Official is not 'significantly' better than OCF)

The usual technique for hypothesis testing is to select \( \delta \) and report a p-value with which we can accept or reject the null hypothesis at the chosen confidence level. However, in this report, we consider a fixed p-value of 0.025 and report the value of \( \delta \) for which the null hypothesis is rejected with a p-value of 0.025.

Rejecting the null hypothesis with a p-value of 0.025 is equivalent to a two-sided 95 per cent confidence interval for \( \Delta BS \) lying entirely below \( \delta \).

This formulation of the null hypothesis implies a default of maintaining the status quo of manual forecasts. We will make the change to automated forecasts only if we have strong evidence against Official being of 'significantly' better quality than OCF. That is, we will automate only if we have strong evidence that automating will not cause a 'significant' degradation of quality. In this context 'significant' means a Brier Score difference of at least \( \delta \).

In this report we construct a “worst case” Scenario based on the upper limit of the confidence intervals for the difference in Brier Score as described in Section 3.1.1. The “worst case” Scenario is intimately related to this null hypothesis formulation.
4. RESULTS

4.1 Australia-wide Results for DailyPoP

Figure 4.1 shows results for the Australia-wide station group. For each lead day from +1 to +7 we show the Brier Scores $BS_{OCF}$ and $BS_{Official}$ as calculated on our sample, their difference $BS_{OCF} - BS_{Official}$, and the 95 per cent confidence interval for that difference as described at Section 3.1.1. The left hand column is for winter 2016 and the second column for summer 2015-16. We display the results for Official and OCF forecasts, calculated based on the filtered and unfiltered data sets as described at Section 2.6. On the upper graphs the y-axis of the plots is reversed so that a better Brier Score is higher on the graph. On the lower graphs, positive differences in Brier Score indicate that Official had a better Brier Score.

In addition, for each season, the dashed horizontal lines in the bottom panel, one for the filtered and one for the unfiltered data, indicate the improvement in Brier Score per lead day based on the Official forecasts as described at Section 3.1.1.

As expected, Figure 4.1 shows that the quality of the forecasts improves with shorter lead-days. This is true for winter and summer, Official and OCF whether using filtered or unfiltered observations.

Looking at all Australian stations together Figure 4.1 shows that in winter 2016 OCF outperformed Official as measured by the Brier Score for forecasts of lead days +2 to +7. This is seen by noting that the confidence intervals for days +2 to +7 lie entirely below 0. On the other hand, Official may well
have outperformed OCF during Summer 2015-16 at all lead days, noted by all confidence interval lying above, or crossing, the zero line.

More detailed examination of regional and sub-regional station groups for lead days +4 to +7 is presented in Sections 4.2 and 4.3. Some exploration of filtered and unfiltered results is shown in Section 4.7.

We summarise some of the information shown in Figure 4.1 in Table 4.1. We provide data from lead days +4 to +7, the primary days of interest for automation, showing a middle and “worst case” scenario for the loss of skill by automating to use OCF instead of Official. These are expressed in terms of lead-days of skill gained or lost as described in Section 3.1.1.

This presentation with colour coding is called a score-card and is a popular way of providing an overview of results when there are numerous sets of data to present.

<table>
<thead>
<tr>
<th>Lead Day</th>
<th>Middle Scenario</th>
<th>“Worst Case” Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winter</td>
<td>-1.8 -1.6 -1.9 -1.8</td>
<td>-0.4 -0.7 -0.8 -0.6</td>
</tr>
<tr>
<td>Summer</td>
<td>0.6 0.2 -0.4 0.6</td>
<td>1.4 0.9 0.3 1.4</td>
</tr>
</tbody>
</table>

Table 4.1 Scenarios of skill that may be lost by automating DailyPoP forecasts. Australia-wide results. See the text for a detailed description. Negative numbers, colour coded blue, indicate skill gained by using OCF.

The middle scenario in Table 4.1 is described at Section 3.1.1. It is based on the observed difference in skill, averaging the filtered and unfiltered results, and is an indication of a likely scenario of the impact of automating by using OCF, assuming no change in skill of OCF and Official from Winter 2016 and Summer 2015-16.

The “worst case” scenario in Table 4.1 is described at Section 3.1.1. It is based on the upper of the filtered and unfiltered confidence estimates for the difference in Brier Score, expressed in days of skill. This “worst case” scenario assists in answering the question “What is the extreme risk of automating by using OCF?” Whether or not the “worst case” scenario overstates the risk of automating depends on where the truth lies compared to the filtered or unfiltered data, and the p-value considered relevant, here set at 2.5 per cent. Furthermore, basing the “worst case” scenario on the upper limit of the confidence intervals assumes that the decision making process has a default of retaining the status quo of manual forecasts. See further discussion in Sections 3.2 and 5.1.

In winter, for days +5 to +7, the middle scenario shown in Table 4.1 is that an improvement of over 1.5 lead days would be gained by automating by using OCF and even the “worst case” scenario is that automating using OCF would provide an improvement of at least 0.6 of a lead day.
On the other hand, as shown in Figure 4.1, automating by using OCF may degrade summer forecasts. For lead days +5 to +7 the middle scenario in Table 4.1 shows that OCF and Official are likely to be within 0.6 of a lead day of skill of each other. The “worst case” scenario is that DailyPoP forecasts may be degraded by up to 1.4 days lead time by automating using OCF.

### 4.2 Winter Regional Results for DailyPoP

<table>
<thead>
<tr>
<th>Lead days</th>
<th>Middle Scenario</th>
<th>“Worst Case” Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4 5 6 7</td>
<td>4 5 6 7</td>
</tr>
<tr>
<td>Australia-wide</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Official better</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tasmania</td>
<td>-1.4 -1.7 -2.4 -3.5</td>
<td>0.7 0.6 -0.1 -0.2</td>
</tr>
<tr>
<td>Victoria</td>
<td>-2.4 -2.1 -2.4 -2.2</td>
<td>1.4 -0.0 0.2 0.8</td>
</tr>
<tr>
<td>NSW &amp; ACT</td>
<td>-1.4 -0.3 -0.3 0.2</td>
<td>1.2 0.0 0.3 0.0</td>
</tr>
<tr>
<td>Queensland</td>
<td>-0.5 -1.0 -1.0 -0.1</td>
<td>0.2 0.7 0.7 0.9</td>
</tr>
<tr>
<td>Northern Territory</td>
<td>0.2 0.6 0.2 0.9</td>
<td>2.7 2.4 1.6 2.3</td>
</tr>
<tr>
<td>Western Australia</td>
<td>-1.6 -1.8 -2.2 -1.3</td>
<td>0.2 -0.4 -0.1 -0.4</td>
</tr>
<tr>
<td>South Australia</td>
<td>-3.0 -2.9 -3.6 -3.7</td>
<td>-0.4 -1.3 -1.1 -1.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Selected sub groups:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Greater Sydney</td>
<td>0.6 -0.2 -0.3 -0.0</td>
</tr>
<tr>
<td>South Coast NSW</td>
<td>1.5 -0.0 -0.3 0.6</td>
</tr>
<tr>
<td></td>
<td>2.6 0.5 0.5 0.7</td>
</tr>
<tr>
<td></td>
<td>3.0 0.4 0.4 1.6</td>
</tr>
</tbody>
</table>

Table 4.2 Scenarios of skill that may be lost by automating DailyPoP forecasts in winter. Regional results. See Section 4.1 for a detailed description. Blue indicates skill being gained by using OCF.

Table 4.2 replicates the winter results of Table 4.1 with extra station groups. More detailed information, comparable to Figure 4.1, is shown in Figure 4.2.

At the Australia-wide level, we noted that automating by using OCF for days +4 to +7 would have generally seen an improvement in the Brier Score of forecasts for winter 2016, with even the “worst case” scenario showing no loss of skill.

Note that for smaller station groups, the decreased sample size reduces confidence in the results and increases the spread between the middle and “worst case” scenarios. However, the Australia-wide result is generally reflected at the regional level. The “worst case” scenario is most dramatic for the Northern Territory, but this is can be attributed to a very small improvement in Brier Score per lead day due to Dry Season climatology (see Figure 4.2) and we therefore consider the risk of automating...
low. For other regions, even the “worst case” scenario is a loss of skill of no more than 0.9 lead days by automating by using OCF. The near zero values for NSW/ACT at lead days +5 to +7 is consistent with that region already relying heavily on OCF to make its Official forecasts for those lead days.

In addition to the regional results, Table 4.2 shows selected sub-groups which have higher risks of automating by using OCF than those shown in the regional or national results. The result for all sub-groups is in Appendix 2. Maps showing the stations in each sub-group are in Appendix 1.

The data from the sub-groups suggests that Greater Sydney and South Coast NSW should be given special consideration if considering automating lead day +4. A further comment on this is made at Section 5.6.
Figure 4.2  DailyPoP Brier Score and Difference in Brier Score by lead day for winter 2016. Regional results. See Section 4.1 for a detailed description.
Figure 4.3  DailyPoP Brier Score and Difference in Brier Score by lead day for summer 2015-16. Regional results. See Section 4.1 for a detailed description.
4.3 Summer Regional Results for DailyPoP

<table>
<thead>
<tr>
<th>Lead days</th>
<th>Middle Scenario</th>
<th>“Worst Case” Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Australia-wide</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>Regions:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tasmania</td>
<td>1.0</td>
<td>0.1</td>
</tr>
<tr>
<td>Victoria</td>
<td>1.0</td>
<td>0.1</td>
</tr>
<tr>
<td>NSW &amp; ACT</td>
<td>0.9</td>
<td>-0.2</td>
</tr>
<tr>
<td>Queensland</td>
<td>-0.4</td>
<td>0.0</td>
</tr>
<tr>
<td>Northern Territory</td>
<td>1.5</td>
<td>1.8</td>
</tr>
<tr>
<td>Western Australia</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>South Australia</td>
<td>0.8</td>
<td>0.0</td>
</tr>
<tr>
<td>Selected sub groups:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northern NT</td>
<td>1.7</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Table 4.3 Scenarios of skill that may be lost by automating DailyPoP forecasts in summer. Regional results. See Section 4.1 for a detailed description. Blue indicates skill being gained by using OCF.

Table 4.3 and Figure 4.3 replicate Table 4.2 and Figure 4.2 but with data for summer.

At the Australia-wide level automating by using OCF for lead days +4 to +7 has a middle scenario of degrading summer forecasts by up to 0.6 lead days with a “worst case” scenario of a loss of skill of up to 1.4 lead days.

The Australia-wide result is generally reflected at the regional level. The Northern Territory has a markedly higher risk of reducing the quality of the forecast compared to the other regions. The minimal risk for NSW/ACT at lead days +5 to +7 is consistent with that region already relying heavily on OCF to make their Official forecasts for those lead days.

In addition to the regional results, Table 4.3 shows a sub-group of Northern Territory stations selected to highlight the signal of the increased risk in the tropical areas. The result for all sub-groups is in Appendix 2. Maps showing the stations in each sub-group are in Appendix 1.

Considering the regional and sub-group results it appears that automating summer forecasts provides the greatest risk in the tropics. The middle scenario for the Northern NT shows around 2 to 3 lead days of skill are likely to be lost by automating using OCF during the summer, with a risk of a greater loss of skill than that.
The tropical wet season should be given special consideration in considering automation, and OCF monitored for improvement.

If a decision is made to automate by using OCF at any lead days during summer, consideration should be given to confining that to an area away from the tropics.

### 4.4 Users with varied sensitivities to rainfall

The Brier Score is a useful summary, but may report an overall improvement whilst a significant subset of users would experience degradation in service. We can use the Relative Economic Value to examine the impact on different users. This section provides some examples which illustrate the information that can be gained from such graphs.

![Relative Economic Value](image)

**Figure 4.4** Relative Economic Value as a function of cost-loss ratio for DailyPoP lead day +5. Official results and OCF results based on unfiltered data.

Figure 4.4 displays results based on unfiltered data. The filtered data (not shown) provides a similar story. The left hand panel shows us that for Winter 2016 at a lead time of +5 days, the OCF forecast would have provided better value to users with a cost/loss ratio of anything between 0.1 and 0.9. These results reassure us that based on the middle scenario and an Australia-wide average, changing to use OCF in winter would impact on almost all users in a positive way regardless of their cost/loss ratio.

The Northern NT Summer 2015-16 graph (right hand panel of Figure 4.4), shows that users with cost/loss ratios below about 0.4 would have been equally well served by OCF as by the Official forecast. Users with costs/loss ratios above 0.4 were better served by the Official forecasts. That is, in Northern NT Summer 2015-16, the Official forecast was better than OCF for users for whom it is relatively expensive to take action against getting wet.

### 4.5 Text Service

Our assessment has been based on the numerical forecasts. However, there are other considerations. Automation using OCF would be likely to have an impact on text and icon services in addition to the quality as measured by the Brier Score. Forecast text and icons are drafted by the GFE software based on values in the forecast grids.
Some forecasters report that they are influenced by considerations of the text forecasts and the message they believe will be received by those using the text forecasts. Without a detailed study on that specific area we cannot comment on the degree to which this impacts on the quality of the numerical forecasts. The data we have on the impact on text and icon services is the Icon Accuracy, or percentage of days for which the observation is matched by whether or not there are rain-drops on the icon forecast.

The icon forecasts have rain drops whenever DailyPoP > 25 per cent based on the service definition in place during the period of this study. We converted the probabilistic forecast to a wet or dry binary forecast based on this threshold, and calculated the Accuracy as per Section 3.1.3, calling it the Icon Accuracy to emphasise the threshold used in the conversion of the probabilistic forecast to a binary one.

The Icon Accuracy would have decreased by using OCF, as shown for Summer 2015-16 in Figure 4.5. Similar results were found for winter forecasts. These results are consistent with forecasters trying to maximise the number of times the icon matched the occurrence and non-occurrence of precipitation.

Optimising Icon Accuracy, based as it is on a 25 per cent threshold for DailyPoP, necessarily involves degrading the quality of the numerical forecasts. For this reason, decisions should not be based on optimising the Icon Accuracy. Consideration should be given to altering the icon service. For example, only showing raindrops when DailyPoP was at least 50 per cent (instead of the current 25 per cent threshold) would remove the necessity of degrading the Chance of Rain forecasts to optimise the Icon Accuracy.

To illustrate the limitations of reducing a probabilistic forecast to a binary forecast, see Figure 4.6. The relative economic value (REV) is described at Section 3.1.2. In Figure 4.6, the REV of the probabilistic DailyPoP forecast is compared to that of using the binary, or deterministic, icon forecast based on Official forecasts of DailyPoP using a 25 per cent threshold. The REV calculation for a binary forecast is a function of the cost/loss ratio $\alpha$, but the contingency table used is the same for all $\alpha$.  

Figure 4.5 Icon Accuracy by lead day for summer 2015-16. Australia-wide results.
The summer 2015-16 probability forecast provided value to users with cost/loss ratios from 0.03 to 0.9 as shown by the REV above zero in Figure 4.6. The binary icon forecast only provided value to users with cost/loss ratios between 0.1 and 0.5. Other than for users with a cost/loss ratio of exactly 0.25, the value of the binary forecast was reduced compared to the value of the probabilistic forecast. A similar result holds for any binary forecast based on a probabilistic forecast. The results are not particular to the icon forecast.

### 4.6 Spring and Autumn Results for DailyPoP

Although no detail is provided here, the results for spring 2015 and autumn 2016 appear to be part way between the winter and summer results presented. If a choice was made to use OCF for some days in winter but not in summer, the results so far suggest a reasonable starting point would be to use it for 5 or 6 months around the winter period. Although we have not explored it, with some subjective input, an option could be to use OCF when the expected synoptic pattern is more typical of winter than of summer.
4.7 DailyPoP1 Results

When considering DailyPoP results, the Chance of Rainfall of at least 0.2mm, we analysed the forecasts based on filtered and unfiltered observations. Consider the winter NSW Greater Sydney results for lead days 1 to 4, shown in the left hand column of Figure 4.7. If the unfiltered results are representative of reality the organisation may be prepared to automate days +1 to +3 based on a middle scenario of no loss of skill and a “worst case” scenario of a loss of skill of approximately one day. However, the organisation may not wish to automate if the filtered results are more representative of reality. In this situation it would be interesting to know if the filtered or unfiltered results were closer to the truth.

Similar analysis to that shown in Sections 4.1 to 4.3 was conducted for DailyPoP1, the Chance of Rainfall exceeding 1 mm. The results were often similar to the DailyPoP results, but without the uncertainty regarding the appropriateness or otherwise of filtering the observations.

The winter NSW Greater Sydney results for DailyPoP1 are shown in the second column of Figure 4.7. Visual comparison to the DailyPoP results increases confidence regarding the Official Brier Score for DailyPoP being better than that of OCF at lead days +1 to +3.
4.8 DailyPoP15 Results

DailyPoP forecasts are linked to forecasts of the chance of exceeding higher precipitation thresholds. If a decision is taken to automate the DailyPoP forecasts for particular lead days, regions and/or seasons, the next question is naturally whether the corresponding OCF guidance could be used to automate the forecasts for higher thresholds of precipitation.

We present results for DailyPoP15, the chance of precipitation exceeding 15 mm, as an indicator of the situation for forecasts using thresholds of precipitation well above 0.2 mm. Regions with too few events to be deemed meaningful are not shown. Other thresholds between 5 mm and 50 mm were assessed and gave similar results to the 15 mm threshold.

Table 4.4 shows the Winter 2016 score card for DailyPoP15 with a middle scenario of largely gaining skill, although there are a couple of exceptions of up to half a lead day of skill lost. The “worst case” scenario for winter, on the whole, shows slightly greater risks in automating than those involved in automating DailyPoP as can be seen when compared to Table 4.2.

<table>
<thead>
<tr>
<th>Lead days</th>
<th>Middle Scenario</th>
<th>“Worst Case” Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Australia-wide</td>
<td>-0.4</td>
<td>-0.7</td>
</tr>
<tr>
<td>Regions:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tasmania</td>
<td>-1.1</td>
<td>-0.9</td>
</tr>
<tr>
<td>Victoria</td>
<td>0.2</td>
<td>-0.6</td>
</tr>
<tr>
<td>NSW &amp; ACT</td>
<td>0.2</td>
<td>-0.7</td>
</tr>
<tr>
<td>Queensland</td>
<td>-0.3</td>
<td>-0.7</td>
</tr>
<tr>
<td>South Australia</td>
<td>0.3</td>
<td>-0.8</td>
</tr>
</tbody>
</table>

Table 4.4 Scenarios of skill that may be lost by automating DailyPoP15 in winter. Regional results. See Section 4.1 for a detailed description. Blue indicates skill being gained by using OCF.

<table>
<thead>
<tr>
<th>Lead days</th>
<th>Middle Scenario</th>
<th>“Worst Case” Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Australia-wide</td>
<td>-0.7</td>
<td>-0.7</td>
</tr>
<tr>
<td>Regions:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tasmania</td>
<td>0.2</td>
<td>-0.2</td>
</tr>
<tr>
<td>Victoria</td>
<td>-0.1</td>
<td>-0.2</td>
</tr>
<tr>
<td>NSW &amp; ACT</td>
<td>-1.0</td>
<td>-1.2</td>
</tr>
<tr>
<td>Queensland</td>
<td>-1.4</td>
<td>-1.0</td>
</tr>
<tr>
<td>Northern Territory</td>
<td>-0.6</td>
<td>-2.3</td>
</tr>
<tr>
<td>Western Australia</td>
<td>0.1</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Table 4.5 Scenarios of skill that may be lost by automating DailyPoP15 in summer. Regional results. See Section 4.1 for a detailed description. Blue indicates skill being gained by using OCF.
The Summer 2015-16 scorecard for DailyPoP15 is shown in Table 4.5. Comparing to Table 4.3, both the middle and ‘worst case’ scenarios indicate smaller risks overall than those associated with automating DailyPoP results.

Any decision regarding automating chance of rain forecasts needs careful consideration of higher rainfall thresholds in addition to consideration of the results for DailyPoP. The results shown here indicate this is particularly the case for winter.

5. FURTHER DISCUSSION

In this section we discuss how the process of verification to support automation decisions can be clarified and formalised and how the power of verification tests can be improved. We discuss some other more subjective factors that may influence a decision regarding automation of numerical forecasts.

5.1 Objective decision making

The most formal way to make a decision is to decide on the criteria in advance. In the language of Section 3.2, that includes specifying a null hypothesis in detail and a significance level at which to reject the null hypothesis. If the decision criteria were agreed upon in advance we could tailor the information provided to match the criteria specified.

In this report we have presented the “worst case” scenario as the upper limit of the 95 per cent confidence intervals. This approach is closely related to taking the null Hypothesis of the form $H_0: \mu \geq \delta$ and reporting $\delta$ (converted in this report to lead days of skill) for which we would reject $H_0$ at the 2.5 per cent significance level.

The “worst case” scenario presented, using the upper limit of the $\Delta BS$ confidence intervals, is based on an expectation that the default decision is to continue the status quo of manual forecasts. If our default is instead to automate we would have a null hypothesis of the form $H_0: \mu < 0$ and would have been concerned with the lower limits of the error bars around $\Delta BS$.

The approach presented in this report, based on a null Hypothesis $H_0: \mu \geq \delta$, risks a situation where OCF may actually be better than Official but we fail to automate by using OCF because of the uncertainty. An approach based on a null Hypothesis of the form $H_0: \mu < 0$ risks adopting OCF because of uncertainty even if it is actually worse than Official. This illustrates the importance of deciding on a null hypothesis prior to assessing the data.

5.2 Improving Confidence in Results

The spread between the middle and “worst case” scenarios indicates the uncertainty in the verification results. A smaller spread corresponds to greater statistical power.

The power of the tests will be increased if the statistics are calculated over larger samples. This comes at the cost of potentially masking interesting smaller scale results. Within this report we have
provided stratification at different spatial scales. To increase the power of the tests, we could calculate results on longer (e.g. six-monthly) time scales, or even combine results for different lead days.

Removing other uncertainty, such as how to interpret the filtered versus unfiltered data would also increase the power of our tests, and the confidence with which we could make recommendations.

Furthermore, examining the results for successive winters and summers will help determine whether any of the results were coincidental rather than ones that would be sustained.

5.3 Weighing risks based on impact

In making a decision we may weigh the risks to various regions or sub-groups within regions and weigh the risks for each season. For example, an organisation may be more risk-averse in regions with larger populations, and during the wet season. Similarly, they may be more risk averse at shorter lead times than at longer lead times. This paper does not address such assessments.

5.4 Relationship to broader service

If text, icon or other services are driven by numerical forecasts, risks to those downstream products need to be assessed to make a fully informed decision. Section 4.5 provides very limited examples of information that could be used to inform considerations of such services.

Furthermore, if the quality of OCF was considered acceptable for DailyPoP but not higher threshold forecasts such as DailyPoP15, or vice-versa, that would add a complication in implementation of automation in a way that provided any efficiencies to the organisation.

5.5 Acceptable Skill Level

An alternative to comparative verification, as conducted here, would be to specify an acceptable quality or skill level which OCF is required to meet. The strategy would be to automate by using OCF once that benchmark is met independent of the ability of manual forecasters. The Brier Score may be an appropriate metric for such a strategy with values selected as a function of lead day and climatology.

5.6 Impacts of, and on, Future Improvements

An underlying assumption of the work presented in this report is that the measured quality of OCF and Official is relevant for making future decisions.

OCF has already been updated since this study (BOM, 2016) and it is a reasonable expectation that OCF quality will improve further over the coming years. The quality of the manual forecast is also likely to improve. In particular, the verification conducted as part of this project may inform improvements of manual forecasts. In addition, the manual forecasts may improve if the connection to derived products, including text and icon forecasts, is removed.

Implementing automated forecasts means we lose the ability to assess forecasters’ skill unless we set up experiments to do so. For example, in NSW/ACT we have very little information about forecasters’
skill for lead days +5 to +7 due to the operational practice of relying heavily on OCF for those forecasts. This means that even if current indications are that automating lead day +4 forecasts, say, might not reduce skill, there may be a desire to delay automating lead day +4 and test for improvements to Official lead day +4 forecasts following exposure of forecasters to verification and/or following automation of lead days +5 to +7. Considerations of such matters may have to be taken into account in a subjective manner and may lead to a more staged implementation of automation than otherwise.
6. CONCLUSION

We have provided an example to present an objective framework for making decisions about automating forecasts, whilst noting some subjective aspects that may be part of the decision.

The results presented for DailyPoP (Chance of Rain) forecasts show that automating by using OCF in winter for days +5, +6 and +7 would likely see a significant improvement in forecast skill, with a ‘worst case’ of losing no more than 1 day of skill in each state/territory other than the Northern Territory where, being the dry season, there is little absolute difference in skill.

Automation of summer DailyPoP forecasts for days +5, +6 and +7 based on OCF would likely see a degradation of more than 2 days of skill in the Northern Territory, and a smaller degradation in forecast skill in many other regions.

Potential negative effects on the skill of forecasts for higher precipitation thresholds should be assessed and considered before extending any automation to those forecasts.

The results are provided only for numerical forecasts of DailyPoP, and do not take into account the impact on the broader service related to those forecasts. Furthermore, they assume that the current performance of Official compared to OCF, as measured for one year, is indicative of future performance.

The process we have presented here can be generalised into a framework which can be applied to assist in operational decisions regarding the suitability of an automation candidate for many forecast parameters.
ACKNOWLEDGEMENTS

Thanks to Gary Weymouth and Philip Riley for discussions regarding the OCF forecasts presented in this report.

Thanks to many people who reviewed early drafts of this report including Beth Ebert, Gary Weymouth, Chun-Hsu Su, Jonty Hall, Neal Moodie, Shaun Cooper and Andrew Marshall.

REFERENCES


APPENDIX 1  STATION GROUPS

The following figures show the stations and station groups introduced in Section 2.1 and used throughout the report.

Figure A.1 Stations used, shown by region. NSW & ACT stands for New South Wales and the Australian Capital Territory.

Figure A.2 Stations used in Tasmania, shown by sub-group. Hobart is within South-eastern TAS.
Figure A.3 Stations used in Victoria, shown by sub-group. Melbourne is within Central VIC.

Figure A.4 Stations used in NSW & ACT, shown by sub-group. Sydney is within Greater Sydney. Canberra is within Ranges NSW/ACT.
Figure A.5 Stations used in South Australia, shown by sub-group. Adelaide is within Central SA.

Figure A.6 Stations used in Western Australia, shown by sub-group. Perth is in South-west WA.
Figure A.7 Stations used in Northern Territory, shown by sub-group. Darwin is in Northern NT.

Figure A.8 Stations used in Queensland, shown by sub-group. Brisbane is in South-east QLD.
### APPENDIX 2  SUB-REGION SCORECARDS

<table>
<thead>
<tr>
<th></th>
<th>Middle Scenario</th>
<th>“Worst Case” Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lead days</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northern TAS</td>
<td>-0.8 -0.7 -1.8 -2.0</td>
<td>0.8 1.1 0.2 -0.8</td>
</tr>
<tr>
<td>South-Eastern TAS</td>
<td>-2.2 -2.6 -2.2 -3.0</td>
<td>0.9 0.7 1.0 -0.9</td>
</tr>
<tr>
<td>South-Western TAS</td>
<td>-1.6 -2.4 -3.2 -4.1</td>
<td>0.2 -0.7 -1.2 -2.0</td>
</tr>
<tr>
<td>Central VIC</td>
<td>-1.7 -2.2 -2.0 -2.0</td>
<td>1.7 0.9 1.2 1.1</td>
</tr>
<tr>
<td>Gippsland</td>
<td>-0.2 -1.3 -0.5  0.4</td>
<td>2.4 1.5 2.2 2.7</td>
</tr>
<tr>
<td>Northern VIC</td>
<td>-3.3 -2.2 -3.0 -2.9</td>
<td>-0.1 0.1 -0.3 0.2</td>
</tr>
<tr>
<td>South-West VIC</td>
<td>-3.2 -2.4 -3.5 -4.2</td>
<td>0.0 -0.0 -0.5 -0.8</td>
</tr>
<tr>
<td>Greater Sydney</td>
<td>0.6 -0.2 -0.3 -0.0</td>
<td>2.6 0.5 0.5 0.7</td>
</tr>
<tr>
<td>Inland Slopes</td>
<td>-3.6  0.0 -0.4 -0.2</td>
<td>0.3 0.9 0.3 0.3</td>
</tr>
<tr>
<td>North coast NSW</td>
<td>-0.5 -0.7 -0.6 -0.6</td>
<td>1.6 0.8 0.7 0.5</td>
</tr>
<tr>
<td>Ranges NSW/ACT</td>
<td>-3.6 -0.2 -0.4  0.4</td>
<td>0.9 0.6 0.3 1.8</td>
</tr>
<tr>
<td>South coast NSW</td>
<td>1.5 -0.0 -0.3  0.6</td>
<td>3.0 0.4 0.4 1.8</td>
</tr>
<tr>
<td>Western NSW</td>
<td>-3.8 -0.3 -0.1 -0.0</td>
<td>0.4 0.3 0.7 0.7</td>
</tr>
</tbody>
</table>

|                    |                 |                       |
| Central East Coast Qld | -1.5 -1.4 -1.2 -0.5 | -0.0 -0.2 -0.4 1.5   |
| Inland Qld          | 0.3  0.0 -0.3  0.3  | 1.0 0.6 0.3 1.2      |
| Northeast Coast Qld | -0.0 -2.0 -2.0 -0.1 | 0.4 -0.6 -0.4 1.4    |
| North-west Qld      | -0.0 -0.6 -1.2 -0.2 | 1.3 0.4 -0.3 1.3     |
| South-east Qld      | -0.5 -1.2 -0.6 -0.2 | 1.5 0.1 0.9 1.1      |
| Northern NT         | 1.3  0.6  0.8  2.0  | 6.2 5.7 5.5 6.1      |
| Southern NT         | -0.4  0.5 -0.0  0.3  | 0.6 1.9 0.9 1.6      |
| Central WA          | -1.5 -1.4 -1.2 -0.4 | -0.2 -0.3 -0.2 1.1   |
| Kimberley           | 0.0  0.0  0.0  0.0  | 0.0 0.0 0.0 0.0      |
| South-west WA       | -1.4 -1.7 -2.4 -1.6 | -0.1 0.0 -0.5 0.5    |
| Upper south-west WA | -2.3 -2.8 -3.1 -1.8 | -0.3 -0.8 -1.0 0.4   |
| Pilbara             | -0.8 -0.5 -0.4 -0.6 | 0.5 0.8 0.8 0.8      |
| Central SA          | -1.6 -1.2 -4.2 -4.6 | -1.0 -0.7 -1.7 1.1   |
| Northern SA         | -0.2 -0.4 -0.7 -0.4 | 1.4 0.8 0.7 1.1      |
| South-east SA       | -3.7 -3.5 -4.3 -4.4 | -0.5 -0.9 -1.5 0.9   |
| Western SA          | -3.1 -3.4 -3.9 -3.6 | -0.5 -0.8 -1.3 0.3   |

Table A.1 Scenarios of skill that may be lost by automating DailyPoP forecasts in winter. Sub-group results.
Blue indicates skill being gained by using OCF. This is an expanded version of Table 4.2.

**Legend**
- Official better: 3
- OCF better: -3
- Too few events in the Kimberley for meaningful results
### Table A.2 Scenarios of skill that may be lost by automating DailyPoP forecasts in summer. Sub-group results.

Blue indicates skill being gained by using OCF. This is an expanded version of Table 4.3.