Seasonal Forecasting for Australia using a Dynamical Model: Improvements in Forecast Skill over the Operational Statistical Model

Andrew N. Charles, Robyn E. Duell, Xiangdong Wang

and Andrew B. Watkins

Bureau of Meteorology, Melbourne, Australia

(Manuscript received April 2014; accepted November 2015)

In 2013 the Bureau of Meteorology transitioned from issuing seasonal climate outlooks based on statistical relationships between sea surface temperature (SST) patterns and Australian rainfall, to a dynamical model-based system, the Predictive Ocean Atmosphere Model for Australia (POAMA). The case for the move to POAMA is threefold: a) when assessed over a common period, POAMA shows higher forecast skill than the previous operational model; b) dynamical models are less susceptible to changes in statistical relationships, whether these occur naturally or through climate change; and c) dynamical models generate physically consistent forecasts of a range of climate phenomena over a number of timescales.

The seasonal climate outlook for rainfall is the most popular product within the Bureau's climate prediction service. It is expressed as a probability of the seasonal (three month mean) rainfall exceeding the long-term median. In this study, independent retrospective forecasts from the statistical model and the dynamical model were assessed over a common period from 1981—2010.

Previous assessments of dynamical model-based forecasts have identified that the forecast probabilities, as generated by a simple count of the frequency of ensemble members exceeding the median, tend to be overconfident. This overconfidence is overcome in practice through a time lagged ensemble strategy. By combining successive burst ensembles initialised on multiple start dates, the spread of forecast outcomes increases and emphatic probabilities are reduced, resulting in outlooks that are more reliable.

A comparative assessment of accuracy and reliability of the new rainfall outlooks with the previous statistical seasonal forecasting system is made using percent consistent, attributes diagrams and the Brier score decomposition. This assessment shows that over the assessment period, the dynamical model-based system is sharper, more reliable and consistently more accurate over a larger spatial domain than the statistical model.

Corresponding author address: Andrew Charles, Bureau of Meteorology, GPO Box 1289, Melbourne VIC 3001
Email: a.charles@bom.gov.au

Introduction

In May 2013 the Australian Bureau of Meteorology (the Bureau) changed its operational seasonal forecasting system from a statistical based system (Drosdowsky and Chambers 2001), to a system based on the Predictive Ocean Atmosphere Model for Australia (POAMA, version M24). POAMA is the Bureau's operational dynamical coupled ocean-atmosphere model-based system (NMOC 2013, Hudson et al. 2013).

The Bureau's seasonal forecast service consists of outlooks for temperature and rainfall for three months ahead. The Bureau's most popular product within this service is the national rainfall outlook, which provides forecasts of the chance of rainfall exceeding the median. This paper compares POAMA rainfall prediction skill with the skill of the previous operational statistical forecast system for Australia. POAMA supports a large number of forecast products, with forecasts also issued for rainfall probability of exceedance, temperature and sea surface temperature (SST).

The seasonal outlooks are used by stakeholders across many sectors for managing risk and decision making. Traditionally, the agricultural sector has been the primary user of for the Bureau's seasonal climate outlook services, for the purposes of making decisions about annual production strategies (Munro 2011). Seasonal rainfall forecasts are also used by the emergency management sector to plan for flood and fire seasons and to inform agencies at all levels of government. Other sectors, including water management, health, energy, tourism, financial services and insurance are increasingly using these services to make planning decisions (Munro 2011).

Several studies have attempted to examine the economic benefits to agriculture of acting on seasonal forecasts to make business decisions. Although it is difficult to quantify monetary gains, economic benefit for the use of seasonal forecasts was found (Meinke and Hochman 2000, Marshall et al. 1996, Wang et al 2009, McIntosh et al. 2005, Asseng et al. 2012). The value was found in using the forecasts to minimise risk, rather than trying to maximise profits based on a single decision. In other words, the benefit was realised over many seasons, not just one.

Seasonal outlooks can be made using statistical or dynamical forecasting systems. Statistical models make probabilistic seasonal forecasts using lagged relationships between rainfall and suitable predictors. In Australia, a lagged relationship exists between Australian rainfall and the SSTs in the Indian and Pacific oceans (Drosdowsky and Chambers 2001). In dynamical models the physics of interactions between and within the ocean, atmosphere, land and cryosphere are used to calculate weather for the coming season.

All seasonal outlooks produced operationally by the Bureau are probabilistic, not deterministic. The physical processes that generate weather variations are by nature chaotic, so predictions of future states of the climate are inherently uncertain. Short-term weather forecasts are so sensitive to small differences in initial conditions that they have a predictability limit of about ten days. However, longer-term forecasts of seasonal statistics, such as three-monthly means, are possible with some accuracy because large-scale processes in the climate system tend to favour certain kinds of weather (Stockdale et al. 1998). Seasonal forecasts aggregate the weather over the prediction period, providing information such as the likelihood of the coming season being drier, or warmer, than normal. While the primary source of seasonal predictability is the El Nino Southern Oscillation (ENSO), other large-scale features of the climate system such as the Indian Ocean Dipole and Southern Annular Mode affect rainfall in different parts of the country at different times of the year (Risbey et al. 2009).

The Bureau's statistical seasonal prediction system, which was in operation between 1998 and 2013, used the first two rotated empirical orthogonal functions (EOFs) calculated from principal component analysis of near-global SST, as predictors of Australian rainfall (Drosdowsky and Chambers 2001). The predictions of the probability of rainfall exceeding the median are made using linear discriminant analysis, with rainfall also represented as nine rotated principal components of one degree resolution gridded rainfall data over Australian land points. The predictors lagged the start of the forecast period by one month and three months. For example, the forecast for winter (June to August) used SST information from February and April. This forecasting system proved to be successful, and was demonstrated to produce reliable probabilistic forecasts in real-time (Fawcett et al. 2010, Fawcett et al. 2005). In 2010 the NINO3.4 index was substituted as the first predictor, due to concerns that a warming trend was causing a spurious loading on the first EOF. This first EOF was understood to capture the predictable variability associated with ENSO, however because of the large spatial extent of ocean warming, the EOF loading was being dominated by an off-equatorial warm signal when in fact the equatorial Pacific was showing cool anomalies. In this paper, the statistical model configuration used for comparative purposes is the same as that used over the period 2010 to 2013, where the predictors were the NINO3.4 index and the second rotated EOF of SST calculated over 1950 to 1999.

Over the past decade it has become clear that the SST EOF time series used for the statistical operational seasonal rainfall outlooks contained trends and that these trends were influencing the outlooks. The skill of many statistical seasonal forecasting schemes is expected to suffer when predictor variables are driven into ranges outside those for which the statistical model was developed. This challenge for statistical models has increased the urgency of developing dynamical model-based outlooks as a means of better dealing with climate risk. A dynamical model-based system, such as POAMA is less affected by a changing climate because it does not rely on historical relationships.

The Bureau’s operational dynamical seasonal prediction system, POAMA, is based on a global atmospheric model coupled to a global ocean model. Seasonal climate forecasts are inherently probabilistic due to uncertainty in the initial conditions, instabilities in the modelled system, and model error. A way of accounting for uncertainties using a dynamical model is to use an “ensemble” approach (Stockdale et al. 1998).

In an ensemble approach, the models are run multiple times with slight differences to obtain a better estimate of the possible spread of future climates, thus generating a number of equally plausible future scenarios ('ensemble members'). Generation of the ensemble should take into account both uncertainties in the initial conditions as well as uncertainties associated with imperfect models. The climate system is chaotic and this means that small, even insignificant, differences in the initial state of the forecast will make ensemble members diverge rapidly with time. This natural uncertainty, together with our imperfect knowledge of the initial conditions, is accounted for by different ensemble members having slightly different initial conditions. Uncertainties in model formulation are usually taken into account by having a multi-model ensemble approach (Doblas-Reyes et al. 2005). Probabilities of certain outcomes, such as chance of exceeding the median, are calculated from the spread of outcomes of ensemble members. This approach effectively takes the model ensemble distribution as a best guess of the probabilities of future states of the system. This provides an adjustment for model biases, because the ensemble distribution for a particular realisation is measured against the model’s own climatological state.

Earlier versions of POAMA-based rainfall forecasts produced probability outlooks that were significantly overconfident, that is, the ensemble spread was too tight, (Langford et al. 2013), precluding its use for operational seasonal forecasts for Australia. This deficiency has been improved in the most recent version of POAMA (M24) as a result of a new ensemble generation scheme which produces an ensemble of perturbed atmosphere, land and ocean states used to initialise the forecasts (Hudson et al. 2013). The new ensemble generation system in POAMA has improved forecast performance for both intra-seasonal and seasonal timescales, primarily through improved reliability of forecasts (Hudson et al. 2013). POAMA (M24) has improved forecast performance at seasonal timescales compared to previous model versions, however as we will demonstrate, overconfidence continues to be exhibited by probability forecasts initalised on a single start date ('burst ensembles').

While statistical models depend on historical relationships between rainfall drivers and rainfall outcomes, dynamical models do not. It has been proposed that changes in climate forcing may project onto the existing modes of variability of the climate system, altering the frequencies and intensities of existing weather regimes (Corti et al. 1999, Palmer and Räisänen 2002). If such theories are correct, climate change may result in changes to the statistical relationships between large scale climate drivers and local climate which are used for statistical modelling, reducing the accuracy of statistically based outlooks. Other studies have examined changes in ENSO due to global warming (Power et al. 2011), which may negatively affect statistical models if the future ENSO differs from the ENSO for which the model was trained.

In this paper we describe the configuration of POAMA used for the operational seasonal outlooks and compare its performance to the previous statistical model (Drosdowsky and Chambers 2001). We also introduce the use of a time-lagged ensemble which further improves forecast reliability and improves the consistency between the forecasts.

Data and Methods

Dynamical Seasonal Forecasting System: POAMA

The operational POAMA system is described in detail in the technical report NMOC (2013). It is composed of an atmospheric model (Colman et al. 2005) with T47 spectral resolution (roughly 2.5 degree grid cells) and 17 vertical levels coupled to an ocean model (Schiller et al. 2002) with 25 vertical levels, 2 degree zonal resolution and 0.5 degree meridional resolution at the equator increasing to 1.5 degree near the poles. This atmospheric resolution is relatively coarse, so coarse that features of the size of Tasmania are not resolved as land. The models are coupled using the Ocean Atmosphere Sea Ice Soil coupler (OASIS: Valcke et al. 2000). A simple bucket model is used for soil moisture (Manabe and Hollaway 1975).

The hindcasts (retrospective forecasts) and real-time forecasts are generated using three slightly different versions of the model, labelled M24A, M24B and M24C. The use of multiple model versions accounts, in part, for uncertainty due to model formulation. M24B applies a flux correction scheme to ocean-atmosphere fluxes described in NMOC (2011). M24C uses the same physics as previous versions of the system, applying two shallow convection schemes consecutively while M24A applies a single shallow convection scheme. ENSO behaves differently in each version of the model and this affects the teleconnection to regional rainfall (Wang et al. 2011).

Ocean initial conditions are generated by PEODAS (POAMA Ensemble Ocean Data Assimilation System: Yin et al. 2011), with land surface and atmospheric initial conditions as described in Hudson et al. (2011). An ensemble of 10 ocean and atmospheric states perturbed around a central analysis is generated using a coupled breeding scheme (Hudson et al. 2013). This results in 11 ensemble members for each model version giving a 33 member ensemble in total.

In this study we examine M24 hindcasts spanning the period 1981—2010. The hindcasts from POAMA-M24 are initialised three times per month, on the 1st, 11th and 21st days of the calendar month. This setup is deliberately similar to the timing of other seasonal-timescale ensemble prediction systems, such as the National Centre for Environmental Prediction's Coupled Forecast System (Saha et al. 2006). The real-time system generates forecasts at a higher frequency, with two initialisations per week.

Climate observing systems have changed substantially over the past three decades, resulting in changes to the data used to initialise POAMA. Retrospective forecasts are initialised using atmospheric and ocean reanalyses generated using data available at the time, while real-time forecasts are generated using the latest available data. Most significant for seasonal forecasts has been the increase in the coverage of ocean observations over the past three decades (Alves et al. 2011). In general it is considered that greater observational coverage leads to better initialisation and more accurate forecasts, though interannual and decadal variability will complicate the observation of this pattern.

Statistical Seasonal Climate Outlook Model

The statistical model data used in this paper consists of a hindcast set and a set of independent forecasts. Initially, hindcasts for the period 1950—1999 were generated using leave-one-out cross validation as reported by Drosdowsky and Chambers (2001), whereby the prediction method was applied using data from all years other than the one being predicted. This cross-validation is used to avoid the artificial skill that can arise from predicting the same data used to train the model. The skill of statistical forecasts cross validated by leave-one-out can suffer from negative artefacts, as noted by Barnston and van den Dool (1993). Such artefacts can be significant in regions of low model skill. To remove this possibility, a hindcast set was generated using leave-three-out cross validation, in which the data from two random years, in addition to the year being predicted, was also left out. This change resulted in an increase of spatially averaged skill. Results in this manuscript are for the leave-three-out hindcast set.

The statistical Seasonal Climate Outlook (SCO) model is trained over the period 1950-1999. During this process a 'leave-three-out' cross validated hindcast set was generated for the training period. Independent forecasts were then produced, using the trained model for the years from 2000-2011. These independent forecasts were generated with the most recently used set of predictors for the operational statistical system, that is, using NINO34 instead of the SST1 EOF pattern.

Analysis Datasets

In this study we assess a period from 1981—2010. For the statistical model this comprises nineteen years of model hindcasts (1981—1999) and eleven years of independent forecasts (2000—2010). The median rainfall from the AWAP dataset (Jones et al. 2009) from the period 1950—1999 was used as a threshold to determine whether seasonal rainfall for the target season was above or below median for the statistical model. This period was chosen because it is the training period for the statistical model. The period 1981 to 2010 was chosen because the dynamical model hindcasts cover this period. The 0.05 degree AWAP rainfall data was first smoothed using a Gaussian filter with a standard deviation of 0.5 degrees, before being bilinearly interpolated to a 1x1 degree grid. The T47 (approximately 2.5 degree) POAMA data was bilinearly interpolated to a 1x1 degree grid before assessment to provide for a consistent comparison. In the case of the forecasts, the probability of exceeding the median was calculated prior to interpolation.

Skill and Reliability Measures

A number of skill measures that can be used to characterise the accuracy of probability forecasts (Wilks 2011, Fawcett 2008). In this study we use percent consistent, attributes diagrams and the Brier score decomposition to evaluate model skill and reliability.

Percent consistent (PC) compares how often the outlook favoured a particular outcome and how often that favoured outcome was subsequently observed. For example, when an outlook indicated a greater than 50% probability of above median rainfall at a grid point and above median rainfall was observed then this is classified as a consistent forecast. If below median rainfall was observed this would be classified as an inconsistent forecast. The calculation of median thresholds uses 'leave-one-out' cross validation to exclude the year being assessed from the statistics.

These probability forecasts are assessed for reliability using an attributes diagram. An attributes diagram plots the forecast probability in discrete bins against the observed frequency of events in each bin, along with a histogram of frequency versus probability. We follow the prescriptions in Broecker and Smith (2007) for locating points on the x-axis at the mean bin probability. To indicate the sampling uncertainty of the observed frequencies, we use the bootstrap technique described in the same paper. In this bootstrap procedure, the observed and forecast time series are resampled 1000 times, the reliability scores for each forecast bin computed for each resampled time series and the 2.5th and 97.5th percentiles of the distribution of scores is plotted. The resampling was applied to the time dimension only, because the data is highly spatially correlated. Perfect reliability is indicated by the points lining up on the 1:1 line. Over-confidence is indicated by the line sloping such that observed frequencies are lower in their deviation from the base rate than forecast probabilities. Reliability could be computed separately for each model grid point as it is expected to vary spatially, however the short time-series length means that the uncertainty in observed frequencies is too great for diagrams generated in this manner to be informative. In practice, the reliability also varies by season. Analysis of reliability for different forecast start months is a useful diagnostic tool, however the overall reliability of the forecasts over all regions and seasons is used to assess whether probabilities are well calibrated, because a reliable forecast system should be reliable uniformly. The frequency distribution of forecast probabilities shows whether the forecasts have a tendency to be emphatic (indicated by large numbers of very high or very low probabilities) or equivocal (large numbers of near 50% forecasts). There is no ’correct’ frequency distribution; instead this information aids the interpretation of the attributes diagram by indicating the relative sharpness of the forecast probability distributions. Sharpness alone is not a measure of forecast quality; in fact forecasts that are sharp but inaccurate are less useful than forecasts that accurately reflect the expected spread (Joliffe 2007).

The Brier score (Murphy 1973), can be decomposed into three components: reliability, resolution and uncertainty. The full score (BS) is given by reliability minus resolution plus uncertainty. The Brier skill score (BSS) is the percentage improvement over a reference forecast, in this case climatology. Reliability (REL) is the error of the forecast probabilities with respect to the observed frequencies: the lower the better. Resolution (RES) is the weighted variance of the observed relative frequencies. The greater the variance, the more the forecasts are able to distinguish events from non-events. The uncertainty is a function only of the climatological rate of the event being forecast. In the case of a two-class system with equal probabilities (such as the probability of exceeding the median), the uncertainty component is 0.25. Values of the BS less than 0.25 are associated with positive skill. For the BSS itself, higher is better with values greater than 0% indicating an improvement over climatological base-rate forecasts.

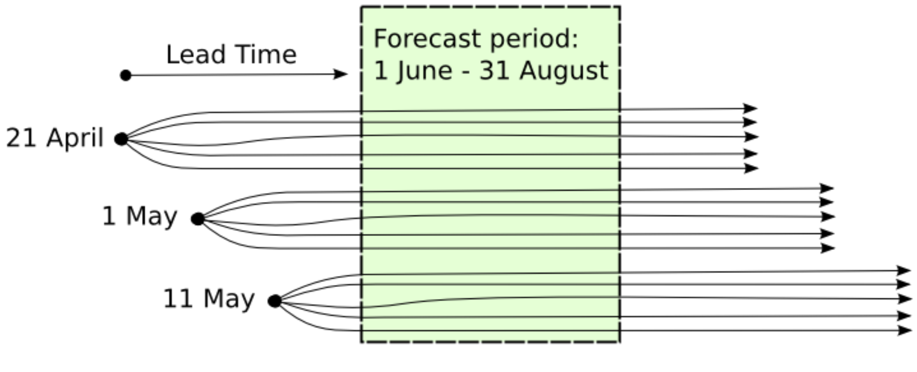
Improving forecast reliability using a lagged ensemble

Overconfidence is a common problem in seasonal prediction for probability forecasts made using coupled ocean-atmosphere dynamical model ensemble frequencies (Mason and Stephenson 2008). Ensembles tend to be under-dispersed and hence produce overly emphatic probability forecasts. This overconfidence means that forecast probabilities are not 'reliable' and therefore do not provide an accurate measure of the risks of a specific outcome occurring. For example, if a forecast system gives a 90% chance of rainfall exceeding median for a given season, then, if the system is to be considered reliable, it is essential that when tested over a large enough sample for all the occasions when the system gave this same chance of exceeding median rainfall for this season, close to 90% of the time rainfall did in fact exceed the median (for a small sample, of course, the statistics would be less stable). If this is not the case then the forecast probabilities given by the model are 'unreliable' as they do not give an accurate forecast of the chance of a specific outcome occurring. Unreliable forecasts are problematic as they can erode forecast value even if the forecast have skill (Vizard et al. 2005). While the multi-model configuration ensemble introduced with the POAMA model version previous to M24 improved the system's reliability (Wang et al. 2011), the resulting rainfall forecasts were still considered insufficiently reliable for operational release. Additional advances in POAMA-M24, including the introduction of a new ensemble generation scheme for characterising uncertainty in initial conditions, also improved forecast reliability for lead times out to a season (Hudson et al. 2013). In order to further improve forecast reliability, the Bureau has also adopted a time-lagged ensemble approach to the generation of the final ensemble for its operational seasonal outlooks. This means that each operational forecast is composed from the output of multiple successive model ensemble runs with different initialisation dates. The lagged approach introduces new sources of difference between ensemble members and increases the dispersion of outcomes.

The choice of the length of lag for the Bureau’s operational system was determined by a balance between improved dispersion (or reliability) and reduced skill due to including model runs with older initial conditions (i.e., longer lead times). Reduced skill may result because using older initial conditions increases the risks of including future states of the atmosphere for which the probability has been reduced to zero in the period since the model was initialised.

It is well known that on average the skill of a seasonal forecast declines with the lead time. In this paper lead time refers to the period of time between the issue time of the forecast and the beginning of the forecast validity period. This is the definition published in the WMO Standardised verification system for long range forecasts (WMO 2006). If the forecast is based on all data up to the beginning of the forecast validity period the lead time is zero. A seasonal forecast issued one month before the beginning of the validity period is said to be of one-month lead.

Figure 1 Lagged ensemble configuration for the forecast period July-August-September (JAS). Size of ensemble at each start date is 33 members such that the final ensemble size is 99.



In this evaluation of the Bureau’s operational lagged system the 99-member ensemble forecast is issued approximately 20 days before the validity period of the forecast. In the hindcast experiment, POAMA-M24 hindcast model runs included in the lagged ensemble are run on the 1st, 11th and 21st of each month, so for example, for a winter forecast with a validity period starting 1 June, the 99-member lagged ensemble is made up from the 33-member ensembles generated on the 21 April, 1 May and 11 May. Thus there is a 40-day lead for the first set of 33 ensemble members, a 30-day lead for the next set of 33 ensemble members and a 20-day lead for the final set. Figure 1**Error! Reference source not found.** shows a schematic of the lagged ensemble strategy.

The current real-time forecast practice is to combine multiple start–dates with an equal-to or shorter-than lead time than the hindcast lagged ensemble examined in this paper. At time of writing, the real-time lagged ensemble contains 165 ensemble members with 33 from each initialisation date, initialised twice per week over a period of approximately 20 days, a difference from the hindcasts where initialisations are separated by ten days over a period of 40 days.

In the next section we demonstrate how the Bureau's operational lagged ensemble system improves the reliability of forecasts and show how model skill is affected.

Model accuracy and the lagged ensemble approach

Figures 2, 3 and 4 show percent consistent for rainfall across the twelve rolling three-month seasons for a hindcast period (1981—2010) of POAMA-M24 for 40-day, 30-day and 20-day lead times for the non-lagged 33-member ensemble. Comparison of these figures reveals that there is only a modest overall decrease in the spatial extent of forecast skill with increased lead time. The figures show that there is large variation across locations and seasons with forecast skill tending to peak in the second half of the year and be overall higher in eastern areas.

Figure 2 Percent consistent rainfall accuracy skill scores of seasonal mean rainfall with a 40-day lead using POAMA-M24 (33-member ensemble) over 1981-2010

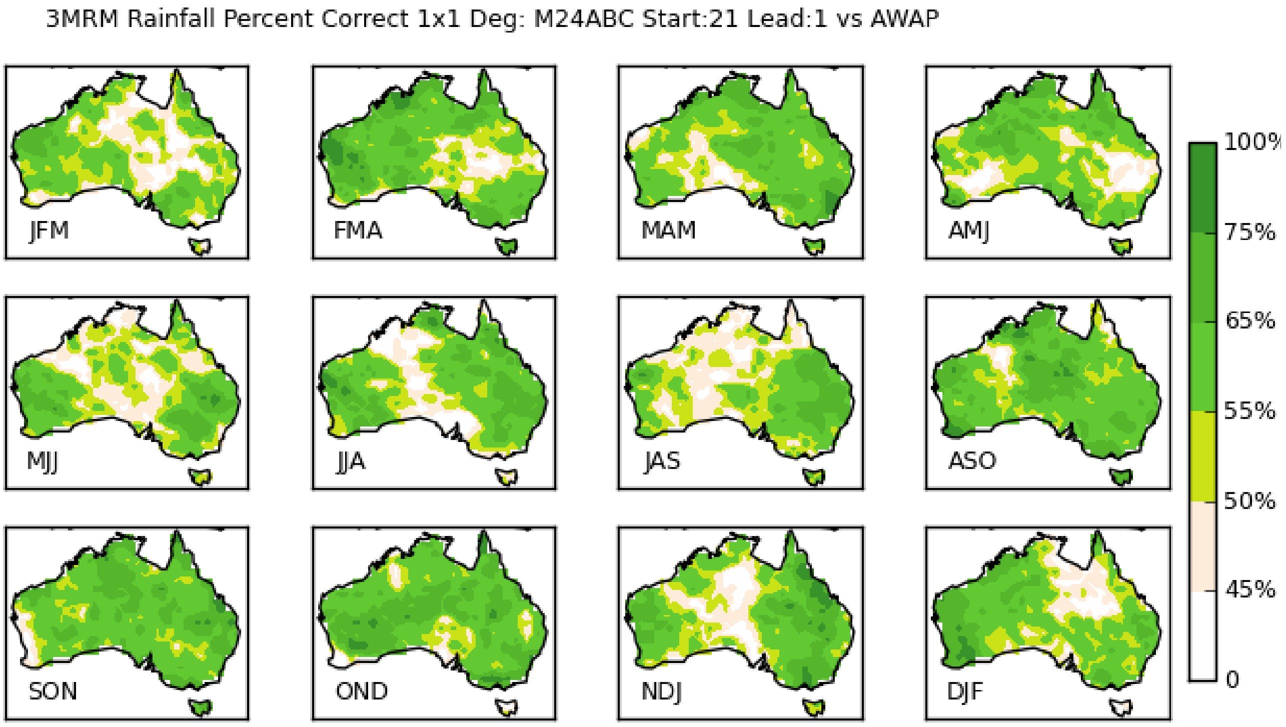

3MRM Rainfall Percent Correct 1x1 Deg: M24ABC Start:21 Lead:1 vs AWAP
JFM
FMA
MAM
AMJ
MJJ
JJA
JAS
ASO
SON
OND
NDJ
DJF
100%
75%
65%
55%
50%
45%
0

Figure 3 As in Figure 2, but for 30 day lead

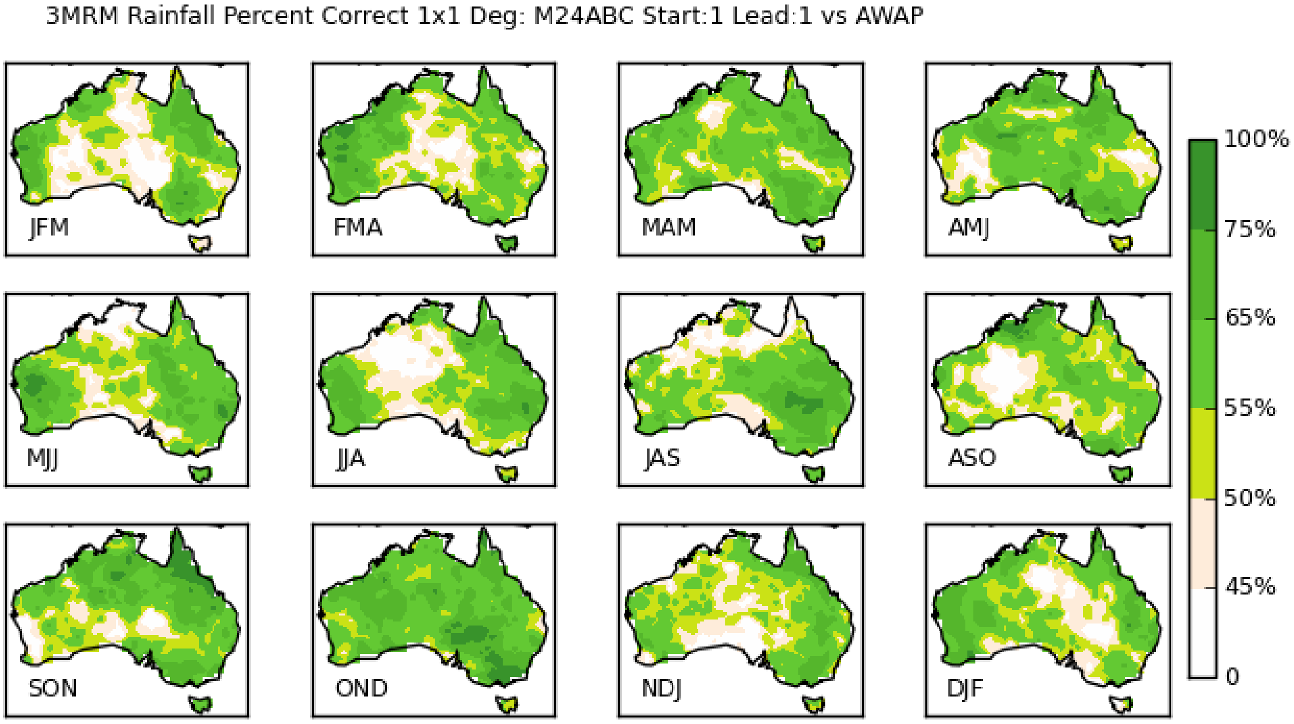

3MRM Rainfall Percent Correct 1x1 Deg: M24ABC Start:1 Lead:1 vs AWAP
JFM
FMA
MAM
AMJ
MJJ
JJA
JAS
ASO
SON
OND
NDJ
DJF
100%
75%
65%
55%
50%
45%
0

Figure 4 As in Figure 2, but for 20 day lead

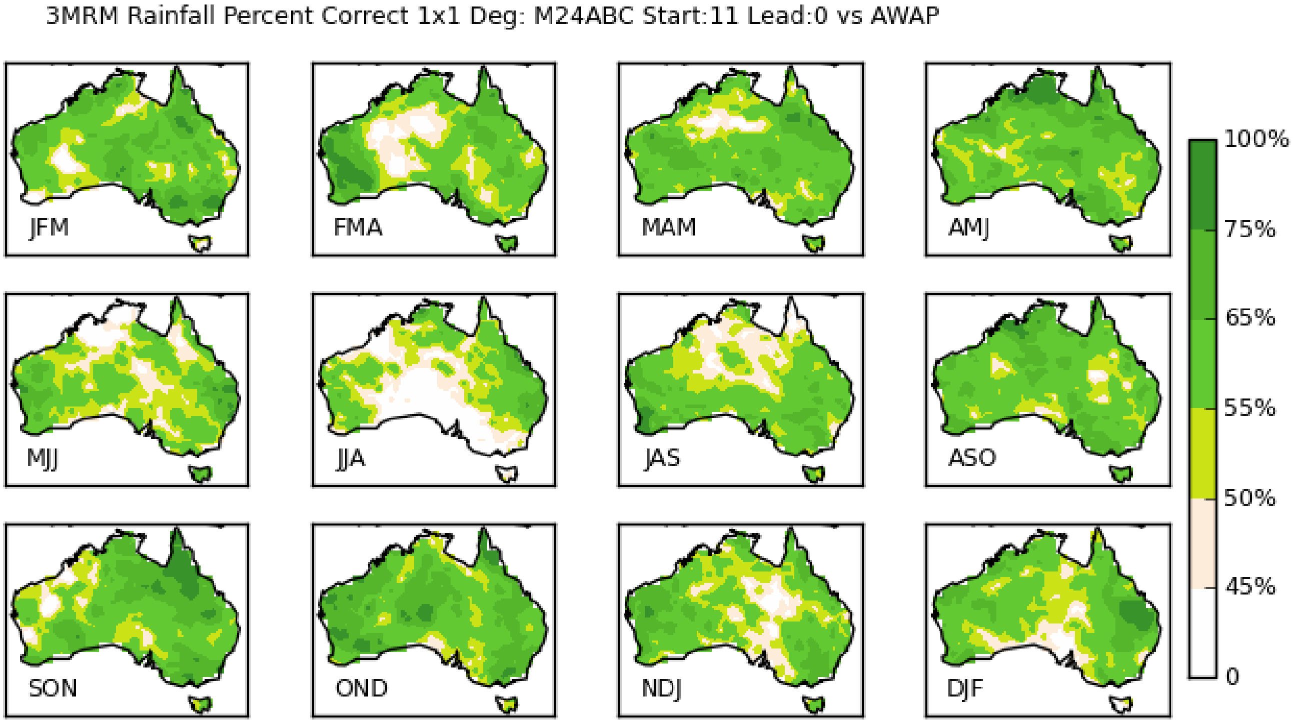

3MRM Rainfall Percent Correct 1x1 Deg: M24ABC Start:11 Lead:0 vs AWAP
JFM
FMA
MAM
AMJ
MJJ
JJA
JAS
ASO
SON
OND
NDJ
DJF
100%
75%
65%
55%
50%
45%
0

Table 1 gives the spatially averaged percent consistent for lead times from 50 days to 0 days, averaged over all land grid points, weighted by the grid point area to take into account the smaller size of poleward grid cells. As expected, on average shorter lead times are generally associated with more accurate forecasts. Because the period of assessment is relatively short the long-term score of the system in real-time might be expected to differ from the score assessed over a limited hindcast period. This uncertainty is estimated using a bootstrap resampling of the temporal dimension of the forecasts and the verifying analysis. Using 1000 such samples with replacement, a distribution for the score is generated and the 2.5th and 97.5th percentiles taken to give a range in which the 'true' score can be found with 95% probability. Limitations of this uncertainty methodology are: it is expected that a small number of years (major ENSO events) contribute significantly to predictability; and resampling the distribution is limited to points inside the assessed period; using this technique the uncertainty for the percent consistent score over 30 years of hindcasts is calculated to be approximately ± 7% (See Table 1). They indicate that for the short time series and limited number of spatial degrees of freedom present, caution should be taken in interpreting very small differences between percent consistent scores, because the probability intervals overlap.

Table 1 Spatially averaged percent consistent scores for each lead time (days) over Australia from POAMA-M24 33-member ensembles (1981-2010). 95% bootstrap probability interval is given in brackets.

Lead days	*Percent Consistent*
50 days	57.4% (49.6-64.9%)
40 days	53.7% (46.1-61.5%)
30 days	54.1% (45.6-62.4%)
20 days	60.1% (52.8-67.5%)
10 days	63.4% (56.9-70.3%)
0 days	61.8% (55.1-68.3%)

A 99-member lagged ensemble is formed from the 20-, 30- and 40-day lead forecasts (see Figure 1**Error! Reference source not found.**). Shorter lead times are omitted because of the practicalities of preparing outlooks and related commentary for public release and because adequate lead time is required for a forecast to make a difference to decision-makers. The accuracy of this lagged ensemble using percent consistent is shown in Figure 5. For this lagged ensemble, the percent accuracy, averaged over Australian land grid points (weighted by area) is 55.9%. Compared to the values of 53.7%, 54.1% and 60.1% for the 40-, 30- and 20-day leads respectively, we see that the shorter lead-time is more accurate by this measure than the lagged-ensemble.

Model reliability and the lagged ensemble approach

Figure 6 shows attributes diagrams for forecasts of the probability of above median rainfall for the period 1981-2010 from the POAMA-M24 hindcast for the 40-day, 30-day and 20-day lead times. Each common start date (or 'burst') ensemble is significantly overconfident. If the forecasts were more reliable, they would sit closer to the 1:1 line and well within the yellow shaded region which defines the region in which points contribute positively to the Brier skill score. Figure 6 (lower-right) shows that the combined 99- member (lagged) ensemble, with an issue-date lead time of 20 days, is clearly more reliable than the individual 33-member ensemble forecasts of which it is comprised.

Skill of above median rainfall forecasts using the statistical and dynamical operational forecasting systems

Figure 7 shows the mapped percent consistent score for the statistical SCO model over the period 1981 to 2010. This skill score shows spatial and seasonal variability, and follows the patterns noted previously by Fawcett et al. (2005, 2010). For ASO, SON and OND the system is able to predict above and below median rainfall categories consistently over most of the continent between 55% and 65% of the time, reflecting the time of year when ENSO is best correlated with Australian rainfall. Skill is patchy and often low across large areas in other seasons, in particular MAM where the model shows no skill over most of the continent. The percent consistent maps for the statistical model show differences, in some seasons, from those calculated over the entire training period (see Figure 8**Error! Reference source not found.**). Independent forecasts using the statistical model for 2000—2009 using only EOF predictors were previously assessed by Fawcett et al who noted the different spatial distribution of skill compared to the hindcast period (Fawcett et al. 2010).

Figure 5 shows the percent consistent score for the POAMA lagged ensemble. The spatial pattern of skill for the lagged ensemble composed from POAMA shows more consistency from season to season than the statistical model (Figure 8). Only in NDJ, OND and arguably JAS is accuracy higher for the statistical model.

Figure 5 Percent consistent of the lagged 99-member POAMA-M24 ensemble, with an issue-date lead time of 20-days over 1981-2010

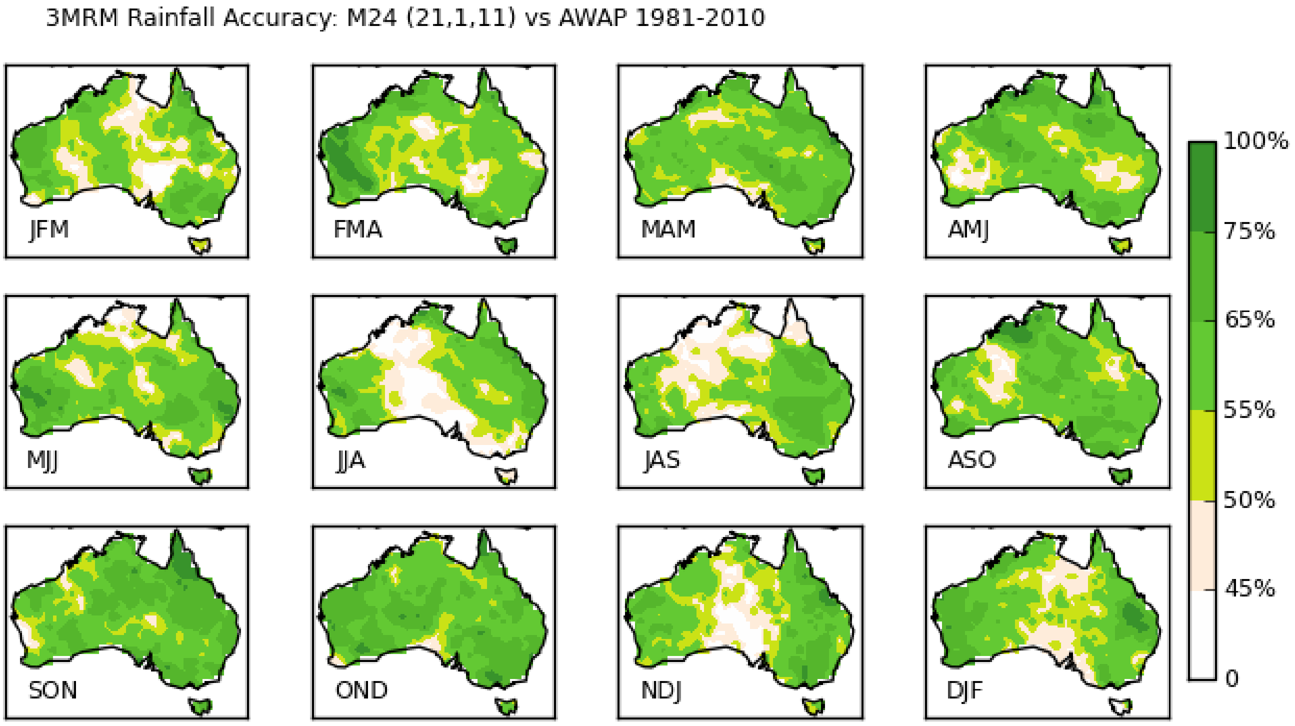


Figure 6 Attributes diagrams of the probability that seasonal mean rainfall is above median (Australian land grid points) using 33-member ensembles from POAMA-M24 for the period 1981-2010. Top left: lead time of 40 days. Top right: lead time of 30 days. Bottom left: lead time of 20 days. Bottom right: Attributes diagram of the lagged 99-member POAMA-M24 ensemble, with an issue-date lead time of 20-days. The blue dashed lines give an indication of the sampling uncertainty at each point using the 2.5% and 97.5% percentiles of a bootstrap resampled dataset. Points in the yellow shaded region contribute positively to the Brier skill score. The inset histogram shows normalised probability density of forecasts in each probability bin

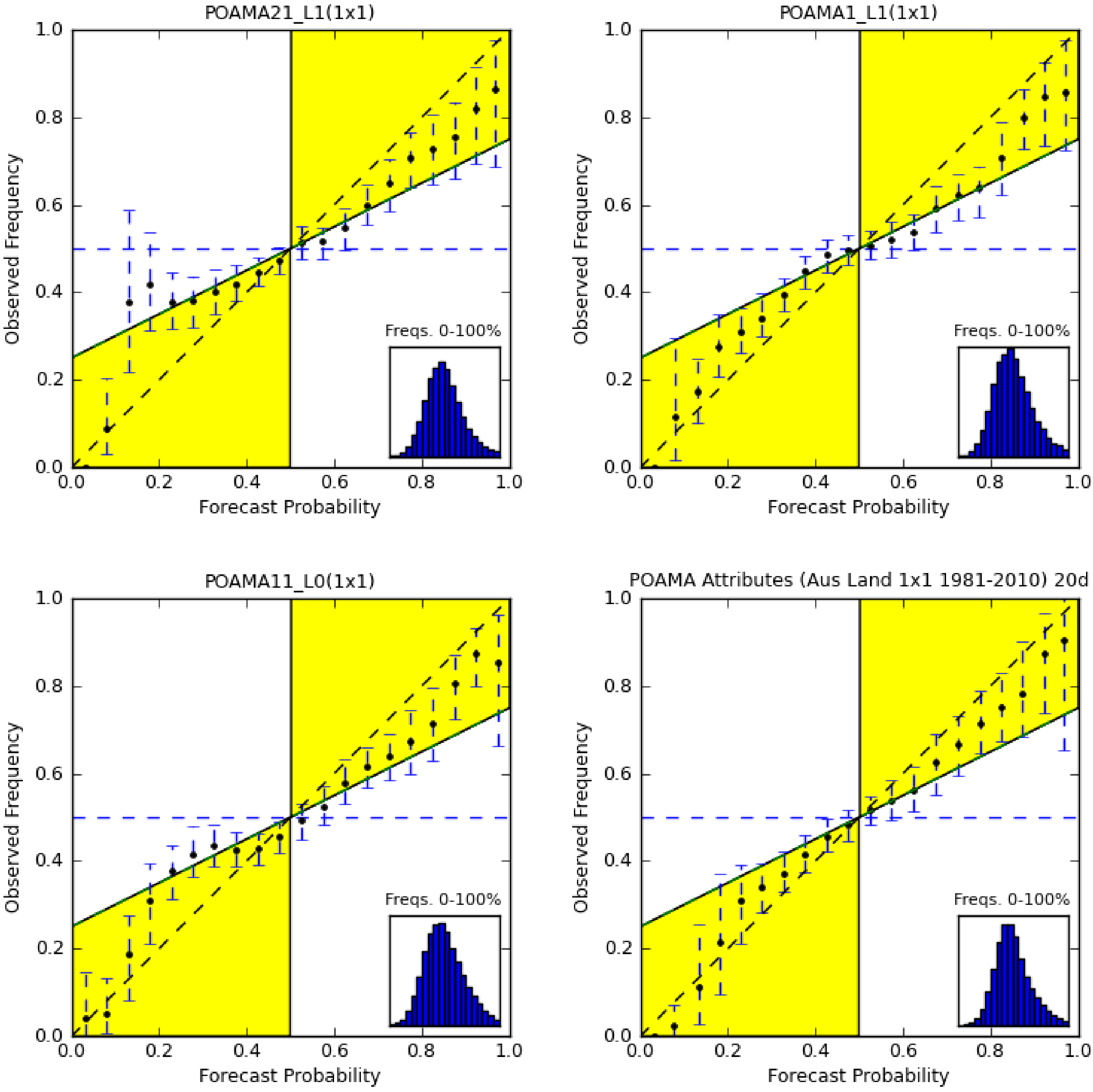


Figure 7 Percent consistent for the statistical seasonal climate outlook model for seasonal rainfall over the period 1981 to 2010

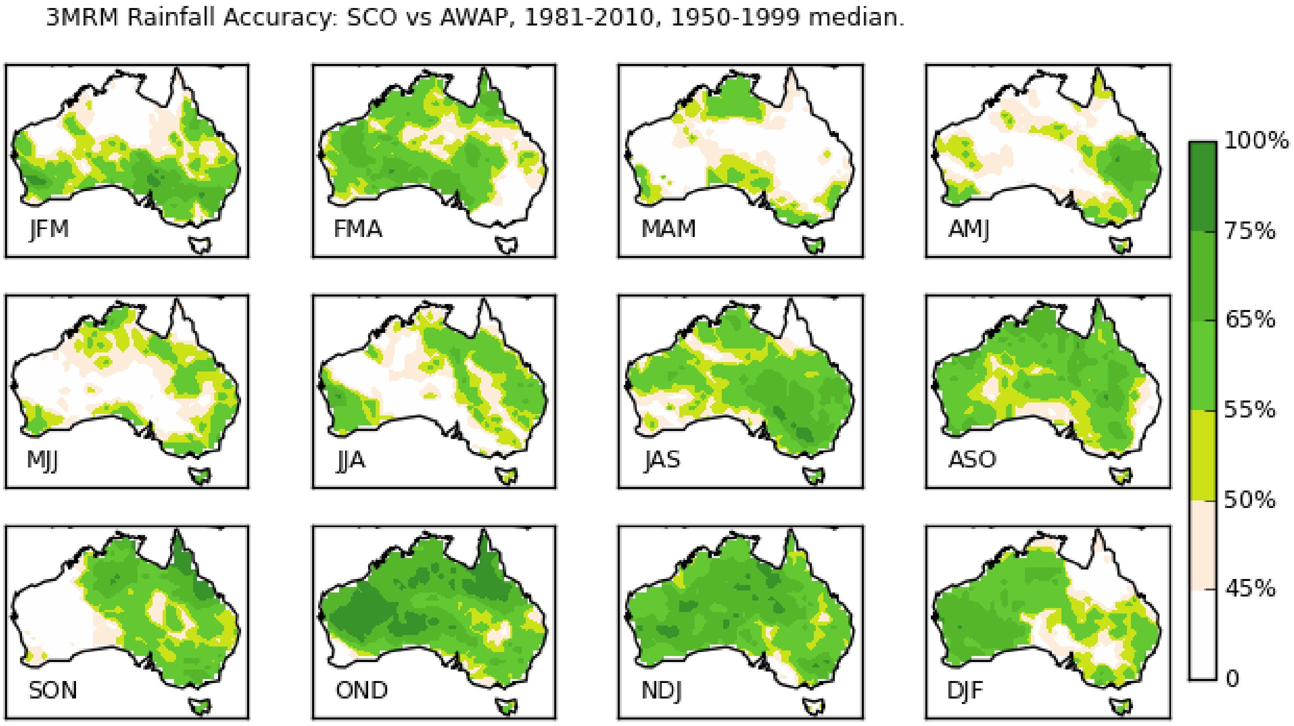

3MRM Rainfall Accuracy: SCO vs AWAP, 1981-2010, 1950-1999 median.
JFM
FMA
MAM
AMJ
MJJ
JJA
JAS
ASO
SON
OND
NDJ
DJF
100%
75%
65%
55%
50%
45%
0

Figure 8 As in Figure 7, over the period 1950 to 1999

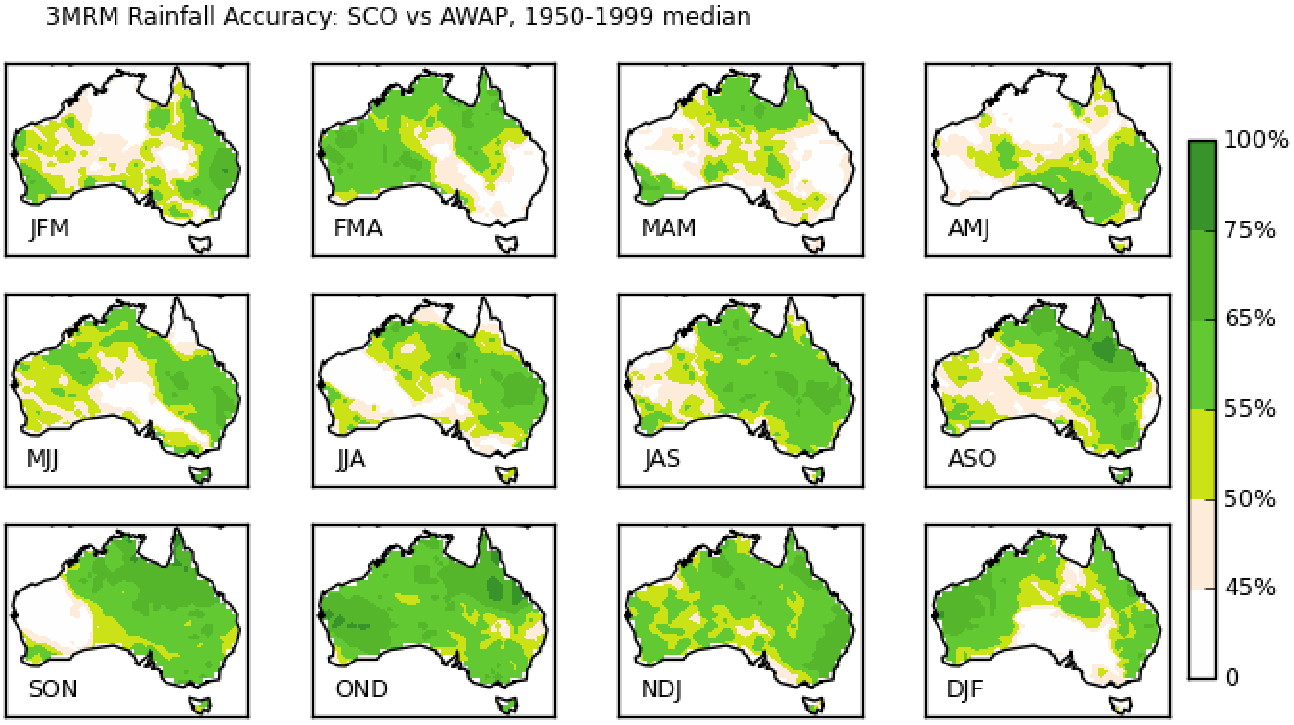


Figure 9 Distribution of percent consistent over all seasons for the Australian region. Bars on the right hand side of the plot (to the right of the black vertical line) denote positive skill. Both systems are biased towards positive skill, with the POAMA lagged ensemble consistently better and having more points for which accuracy is greater than 60%

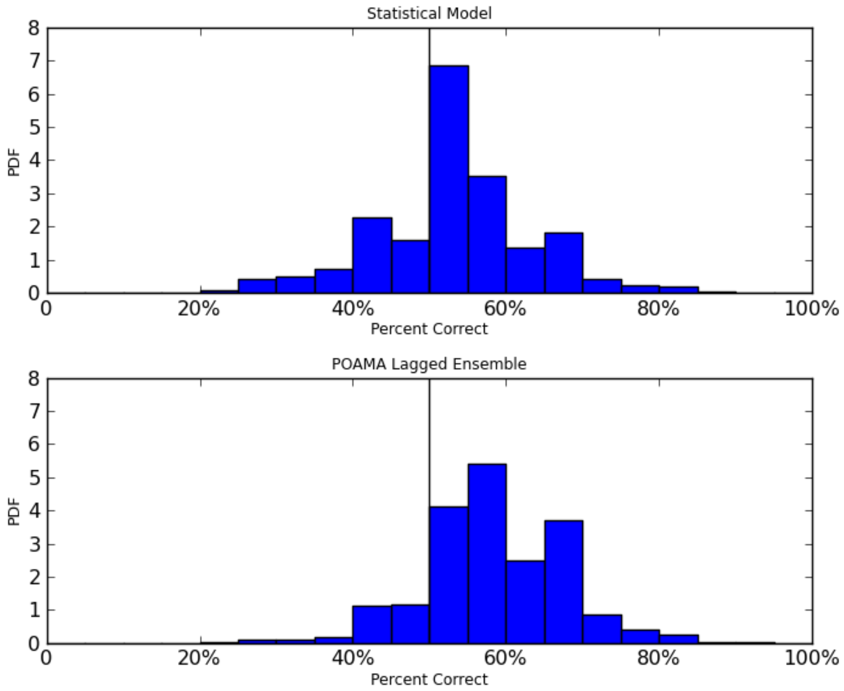


Table 2 Brier skill score (BSS) with reliability (REL) and resolution (RES) components and percent consistent (PC). Scores are calculated over all grid points and seasons over Australia. POAMA-single is for a twenty day lead seasonal forecasts initialised on the eleventh of the month.

System	*Period*	*BSS*	*REL*	*RES*	*PC*
Statistical	1981-2010	0.3%	0.0028	0.0028	51.7% (45.4,58.1)
POAMA-lagged (20 day)	1981-2010	4.4%	0.0017	0.013	55.9% (48.0,64.8)
POAMA-lagged (10 day)	1981-2010	5.2%	0.0015	0.014	54.0% (53.0,69.2)
Statistical	2000-2011	2.7%	0.0022	0.0078	58.3% (47.8,68.7)
POAMA-lagged (20 day)	2000-2011	3.6%	0.0063	0.015	60.1% (44.4,75.9)
POAMA-lagged (10 day)	2000-2011	5.0%	0.0049	0.017	64.0% (47.5,78.7)
Statistical	1950-1979	1.3%	0.00035	0.00338	51.2% (45.5-56.2)
Statistical	1950-1999	0.55%	0.00095	0.00228	50.3% (45.9-54.8)
Statistical	1980-1999	-0.77%	0.00385	0.00173	49.1% (41.9-56.7)
POAMA-single 20 day	1981-2010	2.6%	0.00505	0.0115	60.1% (52.8-67.5%)

The POAMA 20 day lead lagged ensemble, weighted mean percent consistent over Australia is 55.9% with a 95% range of 48.0%—64.8%. The statistical SCO weighted mean percent consistent is 51.7% with a 95% range of 45.4%—58.1%. The better spatial distribution of skill from the dynamical model can be seen clearly by the construction of a histogram of the percent consistent scores for each model aggregated over grid points, as shown in Figure 9. The statistical model (top) shows a distribution only slightly skewed towards positive skill (> 50%). In contrast POAMA shows a clear tendency towards positive skill, with many more points with skill greater than 60%.

Figure 9 Distribution of percent consistent over all seasons for the Australian region. Bars on the right hand side of the plot (to the right of the black vertical line) denote positive skill. Both systems are biased towards positive skill, with the POAMA lagged ensemble consistently better and having more points for which accuracy is greater than 60%

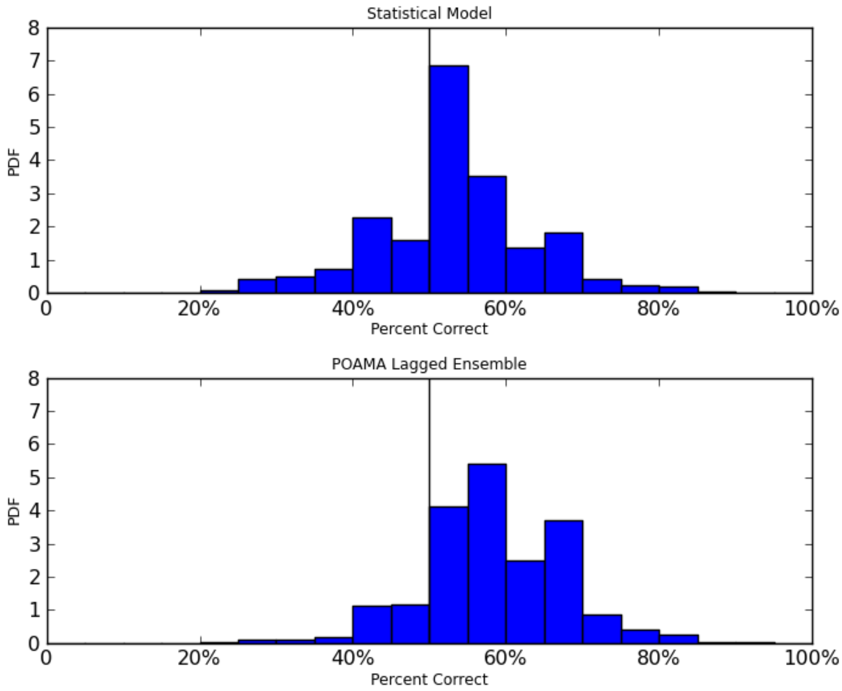


Table 2 presents the Brier skill score calculated for the statistical SCO system over several different periods and for the POAMA lagged ensemble. The uncertainty component is not shown in the table. This score is calculated for the aggregate of all forecasts for all seasons at all grid points over Australia, weighted by the cosine of latitude. The caveat that reliability and resolution will vary by season and by location applies. Nevertheless these measures allow us to quickly compare the forecast systems' overall performance.

Figure 9 Distribution of percent consistent over all seasons for the Australian region. Bars on the right hand side of the plot (to the right of the black vertical line) denote positive skill. Both systems are biased towards positive skill, with the POAMA lagged ensemble consistently better and having more points for which accuracy is greater than 60%

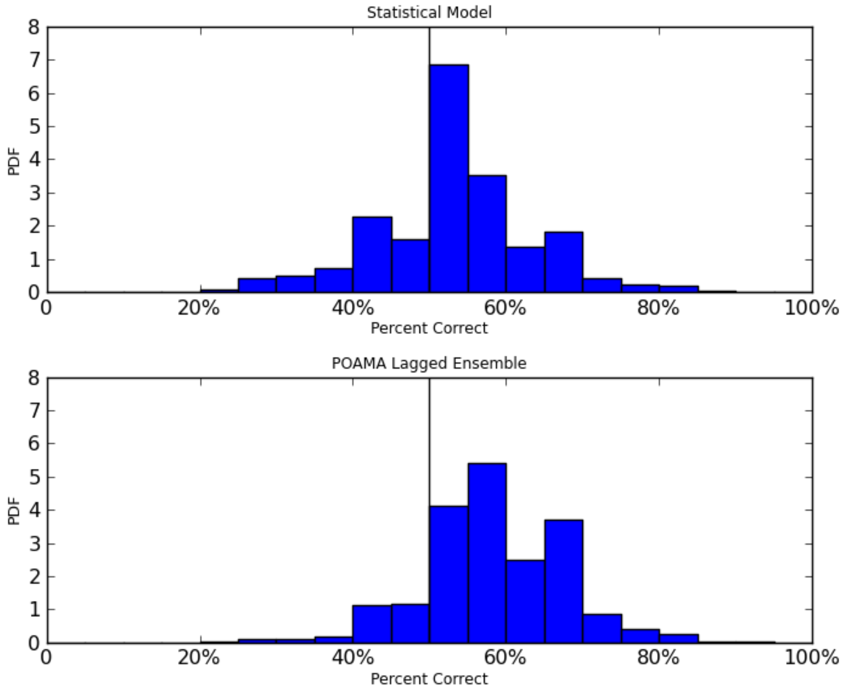


Table 2 shows that the lagged POAMA ensemble outperforms the statistical SCO for the Brier skill score when assessed over the common period. POAMA is marginally more reliable than the statistical SCO and has consistently better resolution. As shown previously, the POAMA lagged ensemble (REL=0.0017) is more reliable than the POAMA single ensemble with a lead of 20 days (REL=0.0051). For operational forecasts it was considered crucial that the forecasts were sufficiently reliable, and that forecasts did not change dramatically from one issue date to another. Both of these limitations were addressed by the use of the lagged ensemble.

Figure 10 shows the attributes of the statistical SCO for the period 1981-2010, while Figure 6 shows the attributes for the lagged POAMA ensemble. Over the period 1981-2010, the statistical model tends towards overconfidence. The overconfidence of the statistical model is tempered by a tendency for forecasts to sit close to 50/50% and is less apparent for the period of real-time operation as shown by Figure 11. In contrast, the POAMA forecasts show greater forecast sharpness, with many more forecasts of probabilities further from the 50% climatological probability. The POAMA outlooks still tend consistently towards overconfidence, but the reliability term of the Brier score decomposition suggests the magnitude of this overconfidence is better than the value for the same period from the statistical model (REL for POAMA-lagged=0.0017, REL for statistical=0.0028). Also given in Table 2 are Brier score decompositions for the statistical model for the periods 1950-1979 and 1980-1999. These results show that the scores can be expected to vary depending on the period chosen, highlighting the non-stationarity of climate predictability.

Figure 10 Attributes diagram for the statistical model for the 1981-2010 period

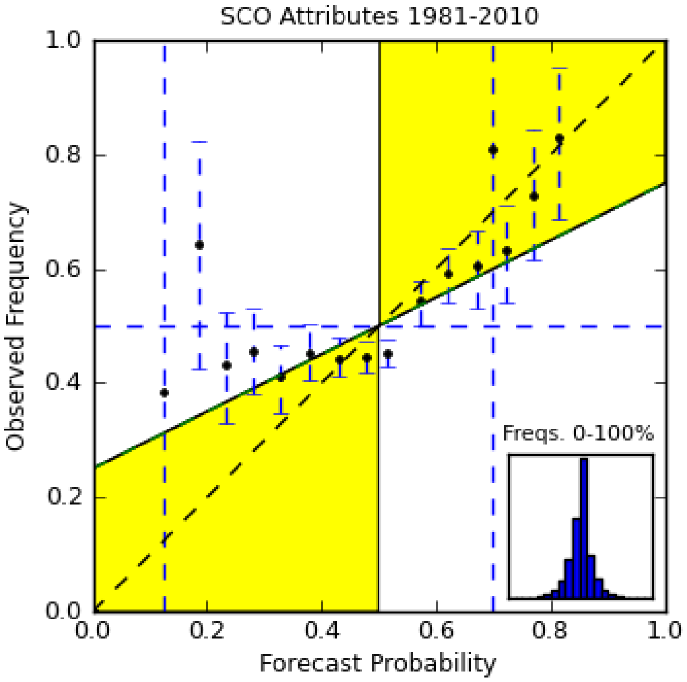
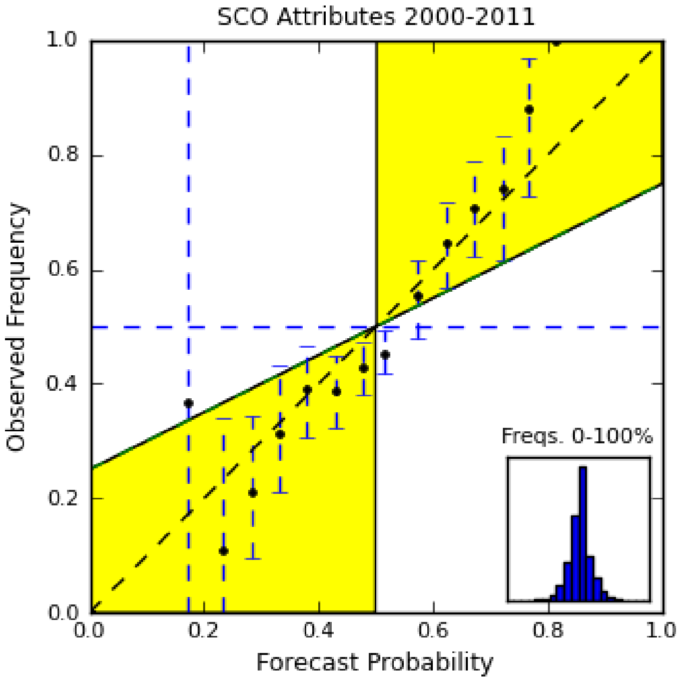


Figure 11 Attributes diagram for the statistical model for the 2000-2011 period



Discussion

In this paper it has been demonstrated that when spatially averaged over Australia, the percent consistent rate for the POAMA lagged ensemble seasonal rainfall outlooks is greater than that expected by chance and greater than the percent consistent rate of the statistical model. It is important to note that the period assessed here for the statistical model is different from the period that was assessed when the model was commissioned; this was done so that we were able to compare a common period with the POAMA hindcasts. Positive skill has previously been demonstrated for the statistical model when assessed over a ten year period of independent forecasts (Fawcett et al. 2005).

Predictability varies decadally (Kirtman and Schopf 1988, DelSole et al. 2014), and can depend on the number and 'flavour' of ENSO events that occur during the period chosen. It is known that ENSO events occur in different forms (Johnson 2013) and that not all ENSO events teleconnect in the same way. There is some evidence that dynamical models can distinguish these differences between ENSO events where statistical models cannot (Wang and Hendon 2007). This highlights the importance of comparing skill measures over uniform periods when comparing different modelling approaches.

In this study we were also able to demonstrate that the lagged ensemble shows improved reliability compared with individual burst ensembles, with most points falling in the region of positive Brier Skill Scores (compared to climatology). Why does the combined lagged ensemble result in a better forecast spread? One explanation is that it samples the range of plausible initial conditions with more breadth than an ensemble based solely on perturbations does. That is, rather than being initialised with states that are tightly clustered around the analysed state, there are three clusters, corresponding to the three included start dates.

The lack of reliability for the statistical model over the hindcast period was unexpected, because it has previously been shown that operational seasonal rainfall forecasts issued over 2000-2009 were quite reliable (Fawcett and Stone 2010). We are able to produce a similar result for the period 2000-2011 (Figure 11). A factor explaining this may be inter-decadal variation in the relationships between ENSO and Australian seasonal climate as already discussed. Another factor is the tendency of cross-validation to negatively influence model skill (Barnston and van den Dool 1993) in the validation period. The set of statistical model seasonal forecasts for 1981-2010 includes a subset of the validation period. Leave-one-out cross validation is known to induce artificial negative effects in model skill assessment (Barnston and van den Dool 1993). While we attempted to reduce this by using leave-three-out cross validation we cannot be sure that negative artefacts do not remain.

Climate risk may be assessed in a historically averaged sense, by using the past distribution of extreme events such as droughts or tropical cyclones to give predictive probabilities of the events in the future. Climate change complicates this approach, because while observed changes in the mean state of the climate system so far has been small, a small change in the mean state can lead to large changes in the frequency and magnitude of extreme events (Rahmstorf and Coumou 2011).

The effect of climate change on weather patterns is likely to be considerably more complex than a simple shift of the existing probability distribution. As an example, a global analysis found a near 50-fold increase in the frequency of extremely hot temperatures during the northern summer, meaning that the historical occurrence now greatly underestimates the risks of extremes (Hansen et al. 2012). It has been proposed that a change in climate forcing projects onto the existing modes of variability of the climate system, altering the frequencies and intensities of existing weather regimes (Palmer 1999, Corti et al 1999). An example of such a mechanism is the prospect that global warming has intensified the hydrological cycle, causing more extreme flooding and droughts (Huntington 2006).

While the empirical relationships between climate predictors and predictands such as rainfall may be robust, in a warming climate, environmental indicators used as predictors can be outside of the range of historical records, meaning that relationships are being assumed for events which do not have an historical analogue. The statistical model showed evidence of climate change significantly impacting the ability of the SST predictors to capture the true state of ENSO. This was reflected in a systematic shift in the first (Pacific Ocean) predictor towards more frequent positive values which was not matched by shifts in the atmosphere.

In general, statistical models cannot reliably account for aspects of climate variability and change that are not represented in the historical record, as statistical forecasting as practised operationally generally depends on the assumption of stationary relationships between predictors and predictands. This also renders such schemes susceptible to periodic changes in these relationships due to decadal timescale variability.

The historical skill analysis in this paper clearly supports the adoption of the operational real-time POAMA system and reveals a substantial improvement in forecast skill. The primary difference between the real-time system and the retrospective forecasts analysed here is the initialisation frequency. At time of writing, POAMA is initialised twice a week, generating an ensemble of 33 forecasts for each initialisation. This enables a larger ensemble with the inclusion of a greater number of forecasts with a shorter lead-time than that studied in hindcast mode in this paper. Another benefit of the lagged ensemble is a smoother transition from one issued forecast to the next, because the forecasts contain information from common model runs. This reduces the risk of dramatic changes between outlooks which may occur when burst ensembles are used.

Another benefit of dynamical model based forecasts is that, because they model the global oceanic and atmospheric circulation, they generate a set of physically consistent future states. These forecasts can be understood in terms of oceanic and atmospheric dynamics and are guaranteed to be self-consistent across a range of forecast timescales and variables.

Conclusions

Over the period 1981—2010, the POAMA lagged ensemble is sharper, more reliable and more consistently accurate than the statistical model it replaced as the basis for the Bureau's seasonal climate outlook service.

Statistical modelling, where the skill is increasingly difficult to gauge due to a changing climate, offers little opportunity to see significant improvements in seasonal forecast accuracy. This study highlights that those statistical models which assume a stationary climate should be viewed cautiously. Future increases in accuracy are likely with dynamical modelling as new science, better modelling techniques, more observations and greater computer power are introduced.

Acknowledgments

The development of POAMA and the work supporting its adoption for operational seasonal forecasting has been supported by countless people and numerous projects run by the Bureau of Meteorology and the Commonwealth Scientific and Industrial Research Organisation (CSIRO). This has occurred through research partnerships such as the Centre for Australian Weather and Climate Research (CAWCR) and via programs such as the Managing Climate Variability Program (MCV), Water Information Research and Development Alliance (WIRADA), the South Eastern Australian Climate Initiative (SEACI), the West Australian Marine Science Initiative (WAMSI), the Great Barrier Reef Marine Park Authority (GBRMPA), the Pacific Adaptation Strategy Assistance Program (PASAP), the Pacific-Australia Climate Change Science and Adaptation planning Program (PACCSAP) and research sponsored by the Grains Research and Development Council (GRDC). Sincere thanks the Luke Garde, Skie Tobin, Debra Hudson, David Jones and two anonymous reviewers whose comments have greatly improved the manuscript.

References

Reference list

Alves, O., Hudson, D., Balmaseda, M., and Shi, L. 2011. *Seasonal and Decadal Prediction*. In *Operational Oceanography in the 21st Century*, Springer, 513-542.

Asseng, S., McIntosh, P. C., Wang, G., and Khimashia, N. 2012. Optimal N Fertiliser Management Based on a Seasonal Forecast. *European Journal of Agronomy*, 38, 66–73.

Barnston, A. G., Li, S., Mason, S. J., DeWitt, D. G., Goddard, L., & Gong, X. 2010. Verification of the First 11 Years of IRI’s Seasonal Climate Forecasts. *J. Appl. Met. Climatol.*, 49, 493–520.

Barnston, A. G., and van den Dool, H. M. 1993. A Degeneracy in Cross-Validated Skill in Regression-based Forecasts. *J. Climate*, 6, 963–977.

Broecker, J., and Smith, L.A. 2007. Increasing the Reliability of Reliability Diagrams. *Weather and Forecasting*, 22, 651−661.

Colman, R, Deschamps, L., Naughton, M., Rikus, L., Sulaiman, A., Puri, K., and Embery, G.. 2005. *BMRC Atmospheric Model (BAM) version 3.0: Comparison with mean climatology.* (No. 108) (p. 32). Bureau of Meteorology Research Centre. Retrieved from http://www.bom.gov.au/bmrc/pubs/researchreports/researchreports.htm.

Corti, S., Molteni, F., and Palmer, T. N. 1999. Signature of recent climate change in frequencies of natural atmospheric circulation regimes. *Nature,* 398, 799–802.

DelSole, T., Yan, X., Dirmeyer, P. A., Fennessy, M., and Altshuler, E. 2014. Changes in Seasonal Predictability due to Global Warming. *J. Climate*, 27, 300–311. doi:10.1175/JCLI-D-13-00026.1

Doblas-Reyes, F. J, Hagedorn, R., and Palmer, T.N. 2005. The Rationale behind the Success of Multi-model Ensembles in Seasonal Forecasting – II. Calibration and Combination. *Tellus A*, 57, 234–52.

Drosdowsky, W., and Chambers, L. E. 2001. Near-global sea surface temperature anomalies as predictors of Australian seasonal rainfall. *J. Climate*, 14, 1677−1687.

Fawcett, R. 2008. Verification Techniques and Simple Theoretical Forecast Models. *Weather and Forecasting*, 23, 1049−1068.

Fawcett, R. J. B. 2008. Verification of the Bureau of Meteorology's seasonal forecasts. *Aust. Met. Mag.* 57, 273−278.

Fawcett, R. J. B., Jones, D. A., and Beard, G. S. 2005. A verification of publicly issued seasonal forecasts issued by the Australian Bureau of Meteorology: 1998-2003. *Aust. Met. Mag*, 54, 1–13.

Fawcett, R. J. B., and Stone, R. C. 2010. A comparison of two seasonal rainfall forecasting systems for Australia. *Australian Meteorological and Oceanographic Journal*, 60, 15−24.

Hansen, J., Sato, M., and Ruedy, R. 2012. Perception of climate change. *Proc. Nat. Acad. Sci.*. doi:10.1073/pnas.1205276109

Hoskins, B. 2013. The potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science. *Quart. J. Royal Met. Soc.*, 139, 573−584. doi:10.1002/qj.1991

Hudson, D., Marshall, A. G., Yin, Y., Alves, O., and Hendon, H. H. 2013. Improving intraseasonal prediction with a new ensemble generation strategy. *Mon. Wea. Rev.*, 141, 4429-4449.

Hudson, D., Alves, O., Hendon, H.H., and Wang, G. 2011. The Impact of Atmospheric Initialisation on Seasonal Prediction of Tropical Pacific SST. *Clim. Dyn.*, 36, 1155–71.

Huntington, T. G. 2006. Evidence for intensification of the global water cycle: Review and synthesis. *J. Hydrology*, 319, 83–95.

Johnson, N. C. 2013. How Many ENSO Flavours Can We Distinguish? *J. Climate*, 26, 4816–4827.

Jolliffe, I. T. 2007. Uncertainty and Inference for Verification Measures. *Weather and Forecasting*, 22, 637–650. doi:10.1175/WAF989.1

Jones, D. A., Wang, W., and Fawcett, R. 2009. High-quality spatial climate data-sets for Australia. *Aust. Met. Oceanogr. J.*, 58, 233−248.

Kirtman, B. P., and Schopf, P. S. 1998. Decadal Variability in ENSO Predictability and Prediction. *J. Climate*, 11, 2804–2822.

Kleeman, R., Moore, A. M., and Smith, N. R. 1995. Assimilation of Subsurface Thermal Data into a Simple Ocean Model for the Initialization of an Intermediate Tropical Coupled Ocean-Atmosphere Forecast Model. *Mon. Wea. Rev.*, 123, 3103–3114.

Langford, S., and Hendon, H. H. 2013. Improving Reliability of Coupled Model Forecasts of Australian Seasonal Rainfall. *Mon. Wea. Rev*, 141, 728–741.

Lim, E.-P., Hendon, H. H., Anderson, D. L. T., Charles, A., and Alves, O. 2011. Dynamical, Statistical–Dynamical, and Multimodel Ensemble Forecasts of Australian Spring Season Rainfall. *Mon. Wea. Rev.*, 139, 958–975.

Lim, E.-P., Hendon, H. H., Hudson, D., Wang, G., and Alves, O. 2009. Dynamical Forecast of Inter–El Niño Variations of Tropical SST and Australian Spring Rainfall. *Mon. Wea. Rev.*, 137, 3796–3810.

Marshall, G. R., Parton, K. A., and Hammer, G. L. 1996. Risk Attitude, Planting Conditions and the Value of Seasonal Forecasts to a dryland wheat grower. *Aust. J. Ag. Resource Econ.*, 40, 211–233.

McIntosh, P. C., Ash, A. J., and Smith, M. S. 2005. From Oceans to Farms: The Value of a Novel Statistical Climate Forecast for Agricultural Management. *J. Climate*, 18, 4287–4302. doi:10.1175/JCLI3515.1

McIntosh, P. C., Pook, M. J., Risbey, J. S., Lisson, S. N., and Rebbeck, M. 2007. Seasonal climate forecasts for agriculture: Towards better understanding and value. *Field Crops Research*, 104, 130–138. doi:10.1016/j.fcr.2007.03.019

Meinke, H., and Hochman, Z. 2000. Using seasonal climate forecasts to manage dryland crops in northern Australia—experiences from the 1997/98 seasons. In *Applications of Seasonal Climate Forecasting in Agricultural and Natural Ecosystems* (pp. 149–165). Springer.

Munro, C. 2011. *Review of the Bureau of Meteorology's capacity to respond to future extreme weather and natural disaster events and to provide seasonal forecasting services*. Australian Government, Department of the Environment. Retrieved from http://www.environment.gov.au/system/files/resources/bc0cc118-a6f2-496c-82fd-0b092c4cc7a5/files/bom-review.pdf

Murphy, A. H. 1986. A New Decomposition of the Brier Score: Formulation and Interpretation. *Mon. Wea. Rev.*, 114, 2671–2673.

NMOC. 2011. *Operational Upgrade to Predictive Ocean Atmosphere Model for Australia Version 2.4 (POAMA-2)* (NMOC Operations Bulletin No. 88). Bureau of Meteorology. Retrieved from http://www.bom.gov.au/australia/charts/bulletins/apob88.pdf

NMOC. 2013. *Operational Upgrade to Predictive Ocean Atmosphere Model for Australia (POAMA-M24)* (NMOC Operations Bulletin No. 96). Bureau of Meteorology. Retrieved from http://www.bom.gov.au/australia/charts/bulletins/apob96.pdf

Palmer, T. N. 1999. A Nonlinear Dynamical Perspective on Climate Prediction. *J. Climate*, 12, 575–591.

Palmer, T. N., and Räisänen, J. 2002. Quantifying the risk of extreme seasonal precipitation events in a changing climate. *Nature*, 415, 512–514.

Power, S. B., and Kociuba, G. 2011. The Impact of Global Warming on the Southern Oscillation Index. *Clim. Dyn.*, 37, 1745–1754.

Rahmstorf, S., and Coumou, D. 2011. Increase of Extreme Events in a Warming World. *Proc. Nat. Acad. Sci.*, 108, 17905–17909.

Risbey, J. S., Pook, M. J., McIntosh, P. C., Wheeler, M. C., and Hendon, H. H. 2009. On the Remote Drivers of Rainfall Variability in Australia. *Mon. Wea. Rev.*, 137, 3233–3253.

Saha, S., and co-authors. 2006. The NCEP Climate Forecast System. *J. Climate*, 19, 3483–3517.

Schiller, A, Godfrey, J. S., McIntosh, P. C., Meyers, G., Smith, N. R., Alves, O., Wang, G., and Fiedler, R. 2002. *A New Version of the Australian Community Ocean Model for Seasonal Climate Prediction*. CSIRO Marine Research. Retrieved from http://ftp.marine.csiro.au/publications/cmrreports/240/title240.pdf

Stockdale, T. N., Anderson, D. L. T., Alves, J. O. S., and Balmaseda, M. A. 1998. Global seasonal rainfall forecasts using a coupled ocean–atmosphere model. *Nature*, 392, 370–373. doi:10.1038/32861

Troccoli, A., Harrison, M., Anderson, D. L. T., and Mason, S. J. 2008. *Seasonal Climate: Forecasting and Managing Risk*. Retrieved from http://dx.doi.org/10.1007/978-1-4020-6992-5

Valcke, S., Terray, L., and Piacentini, A. 2000. *The OASIS Coupler User Guide version 2.4* (Technical Report No. TR/CMGC/00-10) (p. 85). CERFACS.

Van den Honert, R. C., and McAneney, J. 2011. The 2011 Brisbane Floods: Causes, Impacts and Implications. *Water*, 3, 1149–1173.

Vizard, A. L., Anderson, G.A., and D.J. Buckley. 2005. Verification and Value of the Australian Bureau of Meteorology Township Seasonal Rainfall Forecasts in Australia, 1997–2005. *Meteorological Applications*, 12, 343–355.

Wang, E., McIntosh, P., Jiang, Q., and Xu, J. 2009. Quantifying the value of historical climate knowledge and climate forecasts using agricultural systems modelling. *Climatic Change*, 96, 45–61.

Wang, G., and Hendon, H. H. 2007. Sensitivity of Australian Rainfall to Inter–El Niño Variations. *J. Climate*, 20, 4211–4226.

Wang, G., Hudson, D., Yin, Y., Alves, O., Hendon, H., Langford, S., and Tseitkin, F. 2011. POAMA-2 SST skill assessment and beyond. *CAWCR Research Letters*, 6, 40–46.

Wang, G., Kleeman, R., Smith, N., and Tseitkin, F. 2002. The BMRC Coupled General Circulation Model ENSO Forecast System. *Mon. Wea. Rev.*, 130, 975–991.

World Meteorological Organisation. 2006. *Standardised verification system (SVS) for long-range forecasts (LRF) (attachment II. 8), manual on the global data-processing system*. World Meteorological Organization, Geneva.

Yin, Y., Alves, O., and Oke, P. R. 2011. An Ensemble Ocean Data Assimilation System for Seasonal Prediction. *Mon. Wea. Rev.*,139, 786–808.