Australian Government
Bureau of Meteorology

The Centre for Australian Weather and Climate Research
A partnership between CSIRO and the Bureau of Meteorology

CSIRO

Comparison of techniques for the calibration of coupled model forecasts of Murray Darling Basin seasonal mean rainfall

Andrew Charles, Harry Hendon, Q.J. Wang, David Robertson and Eun-Pa Lim

CAWCR Technical Report No. 040

17 January 2011

www.cawcr.gov.au

Comparison of techniques for the calibration of coupled model forecasts of Murray Darling Basin seasonal mean rainfall

Andrew Charles, Harry Hendon, Q.J.Wang, David Robertson and Eun-Pa Lim

The Centre for Australian Weather and Climate Research
- a partnership between CSIRO and the Bureau of Meteorology

CAWCR Technical Report No. 040

17 January 2011

ISSN: 1836-019X

National Library of Australia Cataloguing-in-Publication entry

Author: Andrew Charles, Harry Hendon, Q.J.Wang, David Robertson and Eun-Pa Lim

Title: Comparison of Techniques for the Calibration of Coupled model Forecasts of Murray Darling Basin Seasonal mean Rainfall

ISBN: 978-1-921826-59-7 (PDF/Electronic Resource)

Series: CAWCR technical report; 40

Subjects: Ocean-atmosphere interaction--Australia—Simulation methods.

Sea ice--Computer simulation.

Probability forecasts (Meteorology)--Murray River Watershed (N.S.W.-S. Aust.)

Probability forecasts (Meteorology)--Darling River Watershed (Qld. and N.S.W.)

Rain and rainfall--Measurement--Murray River Watershed (N.S.W.-S. Aust.)

Rain and rainfall--Measurement--Darling River Watershed (Qld. and N.S.W.)

Numerical weather forecasting--Murray River Watershed (N.S.W.-S. Aust.)

Numerical weather forecasting--Darling River Watershed (Qld. and N.S.W.)

Notes: Included bibliography references and index

Other Authors / Contributors: Day, K.A. (Editor)

Dewey Number: 551.6365

Enquiries should be addressed to:

Harry Hendon
Centre for Australian Weather and Climate Research:
A partnership between the Bureau of Meteorology and CSIRO
GPO Box 1289, Melbourne
Victoria 3001, Australia

h.hendon@bom.gov.au

Copyright and Disclaimer

Contents

List of Figures

1. ABSTRACT

Ensemble forecasts of South Eastern Australian rainfall from POAMA 1.5, a coupled ocean-atmosphere dynamical model based seasonal prediction system run experimentally at the Bureau of Meteorology, tend to be under dispersed leading to overconfident probability forecasts. The poor reliability of seasonal forecasts based on dynamical coupled models is a barrier to their adoption as official outlooks by the Bureau of Meteorology.

One approach to correcting this problem is model calibration, in which the probability distribution produced by the model is adjusted in light of available information about its past performance. Several distinct methods for calibrating seasonal rainfall forecasts for South Eastern Australia derived from the POAMA 1.5 ensemble are compared for accuracy and reliability in order to assess their suitability for application to real-time seasonal forecasts.

The calibration methods investigated were: a variance inflation method (IOV); a Bayesian joint probability (BJP) calibration technique; and a singular vector regression technique (SVD) based on co-varying patterns of model and observed rainfall. Calibration was carried out for model grid points in the Murray Darling region.

Assessment was carried out using a mix of standard skill scores widely used in operational forecasting. It was found that the BJP method resulted in the best correction to forecast reliability while IOV improved reliability only modestly and the SVD scheme had a negative impact on reliability. Further study of the application of these methods to real-time forecasts is recommended.

2. INTRODUCTION

Dynamical coupled ocean-atmosphere general circulation models promise to extend the accuracy and lead time with which seasonal rainfall forecasts can be made beyond that of pure statistical schemes. It is expected that in a non-stationary climate purely statistical schemes based on observed relationships between climate indicators and local variables will begin to fail as the climate drifts away from the regime for which the model was tuned. In practice however the best statistical models (for example the current Bureau of Meteorology seasonal climate outlook scheme (Drosdowsky and Chambers, 2001)) still out-perform available dynamical models in the aspect of reliability of probabilistic forecasts.

Reliability, defined as the degree to which the observed frequency of an event coincides with its forecast probability, is an important aspect of probability forecasts. Reliability is essential if probability forecasts are to be used in a quantitative way in risk management or decision making. Reliability does not guarantee useful skill, but forecasts that are not reliable cannot be taken at face value and must be adjusted or 'calibrated', either implicitly as occurs when a verification plot demonstrating overconfidence is published next to a forecast or explicitly by downgrading probabilities that are not justified by model performance. Reliable probabilities are said to be well calibrated.

Analysis of hindcasts of seasonal rainfall from the dynamical model based POAMA system shows that direct ensemble (ensemble relative frequency) forecasts of the probability of above median rainfall in the Murray Darling Basin region (MDB) are overconfident. Ensemble relative frequency means that the probability of the event is assigned as the number of ensemble members in which the event is measured. The assumptions that underpin this method of assigning probabilities are discussed below. Figure 1 shows the reliability diagram over all seasons of the POAMA 1.5b hindcasts for probabilities of above median rainfall prepared in this way for the MDB. The hindcast data used for this analysis are described below.

The reliability diagram (also called an attributes diagram when plotted with a histogram of frequency versus probability) plots the forecast probability in discrete bins against the observed frequency of events in each bin.

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 10th and 90^{th} percentiles of the distribution of scores plotted.

Perfect reliability is indicated by the points lining up on the x=y line. Overconfidence is indicated by the line sloping such that observed frequencies are less (more) than forecast probabilities for above (below) 50%.

The frequency distribution of forecast probabilities (shown inset in the diagrams) 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 reliability diagram.

Reliability could be computed separately for each model grid point as it is expected to vary spatially, however the short time series length means the uncertainty in observed frequencies is too great for diagrams generated in this manner to be informative. Likewise reliability of the direct model output will vary for different seasons, however a good calibration method should be consistently reliable.

In practice the reliability varies by season, with the cooler seasons in which the model has better skill showing better reliability than warmer months in which the model has virtually no predictive skill. Analysis of reliability for different forecast start months is a useful diagnostic tool but the overall reliability of the forecasts over all regions and seasons should be used to assess whether probabilities are well calibrated, because a reliable forecast system should be reliable uniformly. Where no skill exists, reliability is achieved by adjusting the predicted probability distribution to a climatological one.

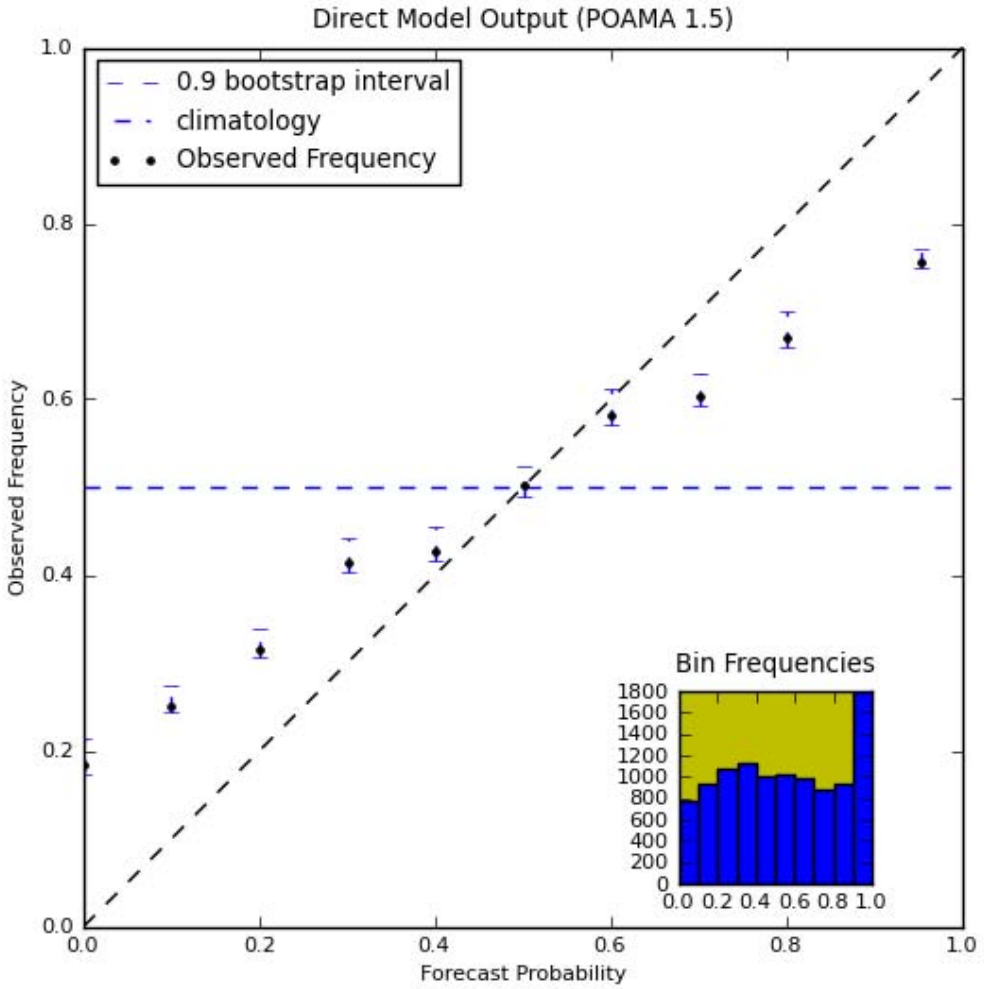

Fig. 1 Reliability of direct model (ensemble relative frequency) output from the POAMA 1.5 hindcasts of the probability of above median rainfall at all grid points in the MDB in the first three forecast months accumulated rainfall in the period 1980-2006, compared with AWAP analysis (Jones et al. 2009). Ten bins for probability categories are used. Blue: climatological probability of above

median rainfall. Bootstrap error bars as described in Broecker and Smith (2007). Inset histogram shows the number of forecasts in each probability bin.

There are several pathways to correcting reliability problems in coupled model forecasts. The first and most obvious is to improve the models by model initialisation. Model improvement work is ongoing but progress is slow, and increases to resolution are limited by computational capacity. The potential of better models in the future does not answer the question of how to extract the most values from imperfect models in the present. A second approach is to combine forecasts from a number of different yet plausible dynamical models. Multi-model combination aims to benefit from a better representation of uncertainty in model physics, model configuration and initialisation strategy. The multi-model approach is widely used in operational weather prediction and extended range weather prediction modelling centres and by organisations that generate seasonal outlooks based on model integrations made publicly available. Model combination is complicated by varying grid resolutions, ensemble sizes, different model skill and mean biases between models, as well as unresolved questions about model weighting. A third approach, model calibration, is the focus of this report. The aim of calibration is to use information about model performance over one period to adjust the forecast for a different period. Calibration methods can be considered to adjust the probability distribution produced by the model by using information about its past performance, with the aim of providing unbiased and reliable probability forecasts.

2.1 Why calibrate dynamical seasonal predictions?

In this section we elaborate on the motivation for model calibration in the context of dynamical seasonal prediction. Due to coarse spatial and temporal resolution and errors in parameterisation of sub-grid processes, dynamical models generate an imperfect projection of the system into the future. This can manifest as systematic mean bias, variability bias (either exaggeration or damping) and as bias in the location, spatial extent and shape of large scale features. A concrete example is the tendency of the ENSO mode in the POAMA model to drift westwards (Zhao and Hendon, 2009).

Ensemble forecasts can be converted into probability forecasts by a number of methods, the simplest of which is to assume the model ensemble is a perfect sample from the space of possible outcomes and to assign probabilities to intervals or percentiles based on the ensemble frequency. This assumes that all uncertainty in the prediction probability density function (PDF) is accounted for by the ensemble spread. We use this relative frequency method to generate the direct model probability forecasts to compare against different calibration schemes. Coupled model ensemble forecasts are typically under-dispersed, leading to probabilities based on ensemble frequencies that are too ‘emphatic’. In this document the term probability is used in the Bayesian sense in which it describes a state of knowledge, in this case the state of our uncertainty about the future given a model forecast and information about its past performance. Consideration of the sources relative frequencies cannot account for all prediction uncertainty with the set of imperfect models currently in operational use at forecasting centres around the world.

The first source of uncertainty is dynamical instability. The atmosphere ocean system is internally unstable such that arbitrarily close initial states will diverge over time. Dynamical ocean-atmosphere models reproduce these instabilities and the tendency for similar initial states to diverge, but not all instabilities in the real system will be captured by models.

The second major source of prediction uncertainty is uncertain initial conditions. Because of the sparseness and imprecision of earth system observations, knowledge of the state of the system with which to initialise models can only be expressed probabilistically. When coupled to the instability noted above, this implies that our knowledge of the evolution of the system in time invariably becomes less certain as lead time from initialisation increases. Ensemble forecasting allows initial condition uncertainty to be estimated and quantified by sampling the space of plausible initial conditions and projecting this sample forward in time.

These first two kinds of error can be described as 'flow dependent' error (see Palmer and Hagedorn (2006) for a detailed account of this concept), because their rates of growth and magnitudes are sensitive to the stability of the point in phase space characterising the flow.

The final source of uncertainty is model error. This includes errors in physical parameterisations and errors due to unresolved processes at the sub-grid scale. Single model ensemble forecasts only capture the components of prediction uncertainty associated with uncertain initial conditions and model-captured instability, and these are only fully captured in the ideal case of an infinite ensemble that uniformly samples initial condition uncertainty. An ensemble of a single model provides no information about the model error component of prediction uncertainty.

Model calibration (also known as 'model error statistics', 'model output statistics' and 'forecast assimilation') is one way to account for model error by adjusting the prediction probability density to be less certain than the ensemble spread indicates. Calibration adjusts the probability distribution produced by the model using information about its past performance. Information about model error is available from hindcasts, also called retrospective forecasts.

A family of model calibration/correction techniques exists with the aim of correcting for systematic errors in the location of patterns or features. While our analysis includes one such technique, in this work we are focussed on the question of whether calibration methods can generate reliable ensemble forecasts. The issue of correcting systematic biases in the location of features is a separate question that is not addressed.

2.2 Qualities of probability forecasts

Reliability, resolution, sharpness and propriety are all relevant properties of probabilistic forecasts (see Jolliffe and Stephenson (2003) for an extensive discussion of these qualities). Reliability is the tendency of the observed frequency of an event to coincide with its forecast probability. Reliability of probability forecasts is essential if they are to be used in a quantitative way in risk management or to make economic decisions. In a simple cost/loss model of determining how much to spend to offset a risk, only a reliable probability allows for an optimal decision to be made (see for example Roulston and Smith (2002)), while forecasts that are not reliable cannot be taken at face value and must be adjusted.

The other main quality of probability forecasts is resolution, which is defined as the frequency with which different observed outcomes follow different forecast categories. Resolution is a property of the model that cannot be improved by simple calibration unless additional information is included (this could be other model variables, or lagged relationships with observations, both of which are better thought of as developing a new statistical dynamical model than simple calibration). Resolution can be degraded by calibration, indeed it is expected that even the best calibration techniques will involve some trade-off in which resolution is exchanged for reliability.

Sharpness refers to the width of the probability distribution, and can be interpreted as the degree of certainty. Forecasts that are over-confident in terms of reliability tend to be too sharp. In an ideal situation sharpness would be reduced for overconfident forecasts with a minimal reduction of resolution.

The quality of propriety of probability forecasts (Murphy and Winkler, 1987) (Jolliffe and Stephenson, 2003) requires that probability forecasts reflect the forecaster's best judgement. 'Improper' probability forecasts are either more emphatic or more equivocal than the forecaster believes is justified. Propriety requires that if forecasters do not believe their probabilities that they should not issue them unchanged. Because they cannot account for all known forecast error, ensemble relative frequencies are not 'proper' predictive probabilities.

3. THE POAMA DYNAMICAL SEASONAL PREDICTION SYSTEM

The coupled ocean-atmosphere dynamical model (General Circulation Model: GCM) POAMA (Predictive Ocean Atmosphere Model for Australia) version 1.5b used in this study is composed of the Bureau of Meteorology Atmospheric Model version 3 (BAM3), coupled every three hours to the Australian Community Ocean Model version 2 with the Ocean Atmosphere Sea Ice Soil (OASIS) coupler (Alves et al. 2003).

The atmospheric model has a spherical harmonic horizontal structure with triangular truncation at wave number 47 (grid cells of roughly 250km by 250km when transformed) and 17 pressure levels. The ocean component has 2 degree zonal resolution with a meridional resolution telescoping from 0.5 degrees near the equator to 1.5 degrees near the poles and 25 vertical levels. This configuration of the model is currently used for sea surface temperature (SST) forecasts for the equatorial Pacific (Wang et al. 2008) and the coral sea (Spillman et al. 2009) at the Australian Bureau of Meteorology.

The hindcasts analysed in this report consist of an ensemble of ten integrations, initialised on the first of each month from 1980 until 2006 using a nudging scheme for atmosphere and land surface initialisation and an optimum interpolation scheme for ocean initialisation. This provides a time series of 27 years in length of the set of forecasts for each start month and lead time.

The atmospheric initial conditions are provided by an Atmosphere and Land Initialization (ALI) scheme (Hudson et al. 2010) which nudges zonal and meridional winds, temperature, and humidity in BAM3 to those of the reanalyses from ERA-40 during 1980-2001 and to the global analyses from the BOM's numerical weather prediction system (GASP) during 2002-2006. The initial conditions produced from ALI are similar to the analyses of ERA-40/GASP but result in less initial forecast shock than if the ERA-40/GASP analyses were directly used as initial conditions. Land surface conditions are initialized indirectly in response to the nudged atmospheric conditions. The ten ensemble members were generated by using atmospheric initial conditions consecutively 6 hours apart going back from the start time.

The ocean data assimilation system provides an estimate of the state of the upper ocean based on the optimum interpolation (OI) of available sub-surface temperature observations (Smith et al. 1991), together with a strong relaxation of the SST to observed analyses.

POAMA's direct rainfall output has some skill at predicting the variations in South Eastern Australian rainfall associated with tropical sea surface temperature at short lead times (Lim et

al. 2009) thought to be a consequence of the model's realistic but exaggerated rainfall response to ENSO.

For this study we extract seasonal mean rainfall as the mean of the first 90 days of each forecast, for the 28 model grid points in the MDB region.

4. CALIBRATION TECHNIQUES

There are a large number of proposed and potential calibration techniques. In this study we compare the performance of three quite different techniques: an inflation of variance (IOV) calibration as described in Johnson and Bowler (2009); a Bayesian conditional probability regression model described in Wang et al. (2009); and a technique designed to correct systematic errors in the spatial location and amplitude of the main modes of variability described in Feddersen et al. (1999). Each of the calibration techniques is applied directly to the coupled model's seasonal mean fields. The dataset used for calibration was the gridded Australian rainfall data described in (Jones et al. 2009) produced under the Australian Water Availability Project (AWAP).

4.1 Inflation of variance

The variance inflation technique used in this work is detailed in Johnson and Bowler (2009). This scheme adjusts the ensemble forecast to meet two conditions: a) that ensemble members have the same variance as observations, and b) that the root-mean-square error of the ensemble mean be equal to the spread of the ensemble. The first condition is designed to achieve the statistical indistinguishability of the first two moments between ensemble members and observations. The second condition is designed to ensure that the ensemble spread accounts for the expected model error. These conditions are achieved by increasing (or decreasing) the perturbations of the ensemble members from the mean while keeping the correlation between model and truth is unchanged (except in the case of a negative correlation between model and truth, in which case the sign of the correlation is reversed). In the case of an under-dispersed forecast this is done by increasing the perterbations of ensemble members from the mean.

Given ensemble mean $\bar{f}$ nd ensemble member perturbations □, adjusted ensemble members g_i are constructed by

$$g_i = \alpha \bar{f} + \beta \epsilon_i .$$

Coefficients α and β are computed as

$$\alpha = \rho \frac{\sigma_x}{\sigma_{\bar{f}}}$$

$$\beta^2 = \left(1 - \rho^2\right) \frac{\sigma_x^2}{\sigma_\epsilon^2}$$

$$\sigma_{\bar{f}}^2$$

with observed standard deviation σ_x, ensemble mean variance , correlation between observations and ensemble mean ρ and time average of ensemble variance σ_ϵ^2. Leave-one-out cross validation is used for the calculation of correlation and standard deviation when constructing a calibrated hindcast set. The time series for each grid point was calibrated independently.

4.2 Bayesian Joint Probability

A Bayesian Joint Probability (BJP) statistical modelling method is described in detail in Wang and Robertson (Wang et al. 2009). BJP is a form of generalised Bayesian regression model in which Markov Chain Monte Carlo (MCMC) sampling is used to estimate transformation and regression parameters. This method is extensible beyond simple model calibration with the inclusion of any number of additional predictors, however in this report we examine solely the simple case of a single model ensemble mean predictor. The ensemble mean time series of seasonal mean rainfall at each POAMA grid point is used as a predictor for seasonal mean rainfall at that grid point, with model parameters computed independently for each grid point and start month.

In this section we give a schematic overview of the technique as applied to this study. It is conceptually similar to the Bayesian conditional probability calibration technique described in Stephenson et al. (2005) and Coelho et al. (2004) but differs in the details of likelihood and parameter uncertainty calculation.

The steps of the BJP calibration are as follows:

1. An extended Box-Cox transform (Yeo and Johnson, 2000) is applied to predictor and predict and data.

2. Transformed data is assumed to be joint-normally distributed and the relationship stationary, so the model parameters are fully specified by the vectors of mean, variance and transformation parameters and a correlation matrix.

3. A transformation is applied to the model parameters (the mean, variance and correlation matrix) such that the re-parameterised correlations to accelerate the MCMC sampling.

4. Model parameters θ are given by Bayes theorem as

$$p(\theta|H,O) = \frac{p(H,O|\theta)}{p(H,O)}p(\theta)$$

with hindcast data H and observations O. $p(\theta|H)$ is the likelihood function, which gives the probability of observing the hindcast-observation series given a set of model parameters.

5. $p(\theta)$ is the prior probability for the model parameters. A uniform prior is used for the Box-Cox transform parameters. More elaborate, diffuse priors are used for the transformed correlation matrix, mean and variance parameters.

6. The distribution of model parameters θ is determined by solving Bayes theorem using MCMC sampling.

7. The posterior probability of the Box-Cox transformed forecast variable is given by a normal distribution, or equivalently a linear regression using the sampled mean and covariance vectors.

8. To generate an ensemble forecast the posterior probability distribution is sampled and an inverse transform is applied to the ensemble forecast.

In the single predictor case examined in this study the model is closely related to a linear regression model but enhanced by accounting for parameter uncertainty and with the potential to detect nonlinear predictor-predictand relationships through the distribtion transformation parameters. Only the ensemble mean is used, so no information about the model spread is included in the calibration, in other words all prediction uncertainty is due to accounting for model error and parameter uncertainty.

In constructing a calibrated hindcast set leave-one-out cross validation was used. The time series of each grid point, for each three month season, was calibrated independently.

4.3 Singular Value Decomposition Analysis Calibration

Finally a technique designed to correct systematic errors in the spatial location and amplitude of the main modes of variability was examined. This method is described in Feddersen et al. (1999), and its application to POAMA 1.5 discussed in Lim et al. (2011). A singular value decomposition analysis (SVDA) of the covariance matrix of the POAMA hindcast ensemble and the National Climate Centre's 'Barnes' gridded analysis (Jones and Weymouth, 1997) and was employed to determine the spatial patterns with the most temporal covariance. This results in a set of model patterns with corresponding (corrected) observed patterns for each forecast season and start month. Although all other techniques were built using the AWAP analysis, at the coarse spatial and temporal resolution used in this study this is not considered to affect the results, as in the observation rich Murray Darling region there is very little difference between the two datasets.

The calibrated forecast is reconstructed using the first five observed patterns with weights given by the projection of the un-calibrated forecast onto the corresponding model patterns. This method is also referred to in the literature as maximum covariance analysis (MCA) Stephenson et al. (2005).

Leave-one-out cross validation is used such that the SVD patterns used for each hindcast year are based on the statistical relationship between observations and the POAMA hindcasts in all other years. Means and standard deviations from the independent period are used at all stages of the calculation.

5. RESULTS

5.1 Inspection of time series

In this section we examine basic properties of the raw and calibrated time series at high and low skill points. Inspecting the time series at grid points of 'high' and 'low' skill gives insight into the characteristics of each calibration method. A perfect calibration method is expected to leave the high skill prediction relatively unchanged (although possibly moderated by sampling uncertainty in the high correlation). On the other hand a perfect calibration method is expected to replace a low skill prediction with a climatological prediction.

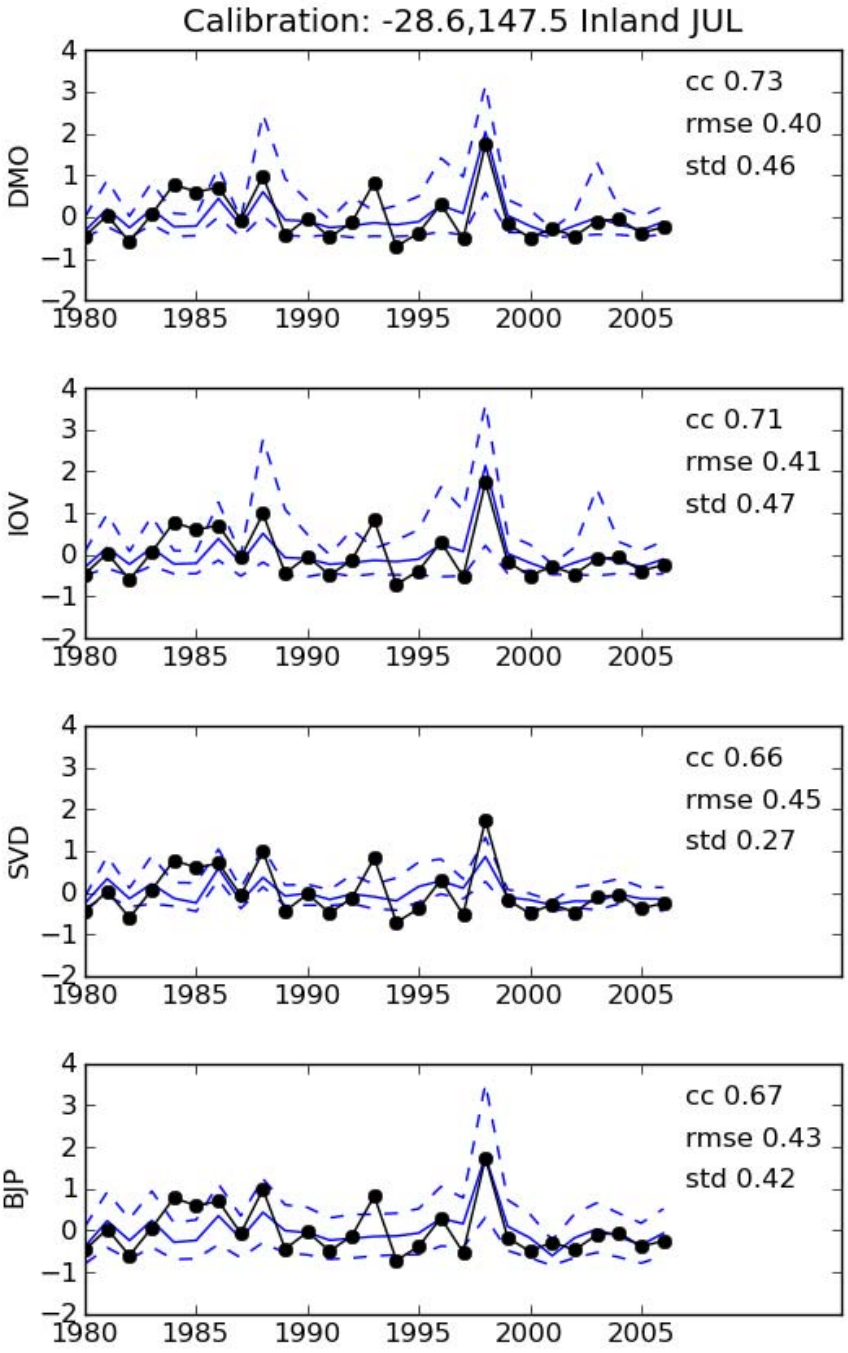
Calibration: -28.6,147.5 Inland JUL
DMO
cc 0.73
rmse 0.40
std 0.46
IOV
cc 0.71
rmse 0.41
std 0.47
SVD
cc 0.66
rmse 0.45
std 0.27
BJP
cc 0.67
rmse 0.43
std 0.42
4
3
2
1
0
−1
−2
1980
1985
1990
1995
2000
2005

Fig. 2 Direct model and calibrated time series of seasonal mean rainfall anomalies starting in July for a point that shows high hindcast correlation. In order from top to bottom: Direct coupled model anomalies, variance inflation, SVD and BJP. Blue lines: tenth percentile (dashed), mean and ninetieth percentile (dashed) of the ensemble at each time step. Anomalies are with respect to each model's climatological mean. cc: Correlation coefficient, rmse: root mean square error between ensemble mean and observations, std: standard deviation of ensemble mean.

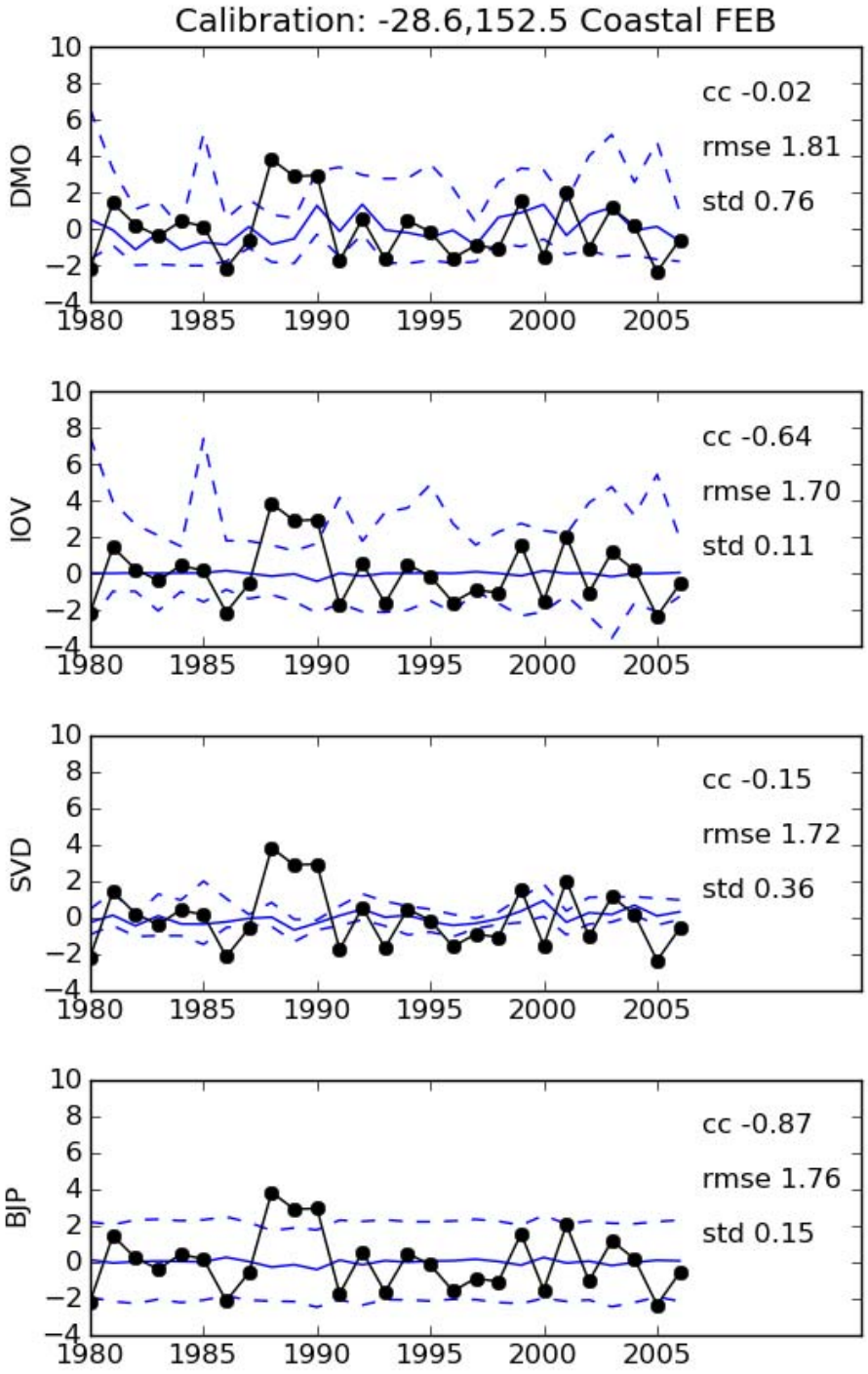

Fig. 3 Direct model and calibrated time series for a point that shows low hindcast correlation. Legend as for fig. 2. Large negative correlations for the BJP and IOV time series is shown to illustrate a cross-validation artifact.

As a high skill point we choose an inland point for the season June-July-August, where the correlation between model and observations is 0.73. Figure 2 shows the time series for the three calibration methods for the season. Applying variance inflation preserves the main phase characteristics of the time series while increasing the distance between highest and lowest ensemble members. The ensemble mean is largely unchanged after variance inflation. The SVD calibration contracts the ensemble spread sharply, even after scaling the normalised svd time series by the observed standard deviation.

The main features of the direct model time series including the peak in 1997 are preserved. The BJP calibration slightly reduces the variability in the ensemble mean, and slightly smooths the year to year variability of the ensemble spread while preserving the main phase characteristics of the un-calibrated ensemble mean. The BJP calibration also increases the spread of the ensemble, capturing more of the observed values between the 10th and 90th percentiles. The RMS errors shown on the figure are of a comparable magnitude between the various methods to within sampling variability. See fig. 12 for a comparison of the RMSE of each method.

For a 'low skill' point we chose a grid point on the New South Wales coast for forecasts starting in February, in which the hindcast correlation is effectively zero. Figure 3 shows the time series of direct model and calibrated output for this 'low skill' grid point.

The variance inflation increases the ensemble spread such that more points are 'captured' in its envelope, and significantly smooths out the variance of the ensemble mean. The SVD calibration again greatly reduces the ensemble spread, which in this case is undesirable. The BJP calibration in this case replaces every forecast with a near climatological PDF. The high negative correlation shown on the plot is an artefact of cross validation and is discussed further below.

6. VALIDATION OF CALIBRATED HINDCASTS

We validate the calibration methods against the direct model output using a number of measures and scores, because no single score can capture all the attributes of good probability forecasts. Some scores are widely used for scoring operational forecasts while others are less widely known. There is inevitably some redundancy between the skill measures we have chosen (for example hit rate and anomaly correlation are both sensitive to phase errors but insensitive to magnitude errors). The gridded Australian rainfall data described in (Jones et al. 2009) produced under the Australian Water Availability Project (AWAP) are used as the verifying analysis.

For category probability type scores (hit rate, reliability) unless otherwise noted we are scoring for the probability of above median rainfall, where the verifying event is observed rainfall for the three month season falling above the climatological median. For direct scores we are verifying against the total seasonal rainfall over the three months from the start date. For most skill measures, the direct model output, SVD and inflation of variance results have their mean subtracted and the observed mean added. This is a common procedure that removes the model's systematic mean bias, and in this work is performed in a leave-one-out cross-validated manner such that for each point in the time series the mean of all other times is used for bias correction. This procedure is not performed for the BJP results because systematic mean biases are corrected by the technique itself.

6.1 Anomaly Correlation Coefficient

Anomaly Correlation Coefficient (ACC) indicates the strength of the linear relationship between two time series. Correlation coefficient can be sensitive to a small number of large events, and is somewhat noisy in that a large sampling error is expected for short time series. Despite these issues, correlation is an excellent indicator of model skill, because it is insensitive to climatological differences between the model and observations, picking out where the two time series have a common signal. Even where the correlation is less than traditional significance thresholds for independent time series, the spatial coherence of correlation patterns can be used to infer that a real signal is present. Correlation is not sensitive to the absolute magnitude of co-variation and so should be interpreted alongside an error statistic. Because we compute correlation for the ensemble mean, it is insensitive to ensemble spread. It is included in this study to assess whether any of the techniques degrade the predictive skill of the raw model.

For forecast time series f with mean $\bar{f}$ and observed time series o with mean $\bar{O}$ and standard deviation σ_o the anomaly correlation is computed as $ACC =$

$$ACC = \frac{(f - \bar{f})(o - \bar{o})}{\sigma_f \sigma_o}. \quad (1)$$

As noted in (Barnston and van den Dool, 1993), the leave-one-out cross validation technique results in a purely artificial anti-correlated signal appearing in the calibrated output. In the absence of a significant relationship between the un-calibrated model and the real system, this artificial anti-correlation dominates the calibration and can result in a calibrated forecast perfectly out of phase with the observations (with small magnitude). This occurs most noticeably in the BJP results. Barnston and van den Dool (1993) note that this degeneracy is only noticeable when the correlation between predictor and predictand is well below typical significance thresholds.

This is demonstrated in Fig. 3 which shows the time series for a grid point for which the model has no discernable predictive skill (anomaly correlation of -0.02). It can be seen by inspection that the BJP calibration is the better prediction of the three time series because it reverts to a climatological PDF. The high anti-correlation of -0.87 of the BJP ensemble mean with the observed time series is due solely to the small amplitude fluctuations caused by cross-validation. The other calibration methods also show high negative correlations for this case.

These spurious large negative correlations should be ignored when interpreting the results. The simplest way to achieve this, suggested by Barnston and van den Dool (1993) and adopted in this study is to treat all negative correlations as zero when plotting, and when computing averages. We have no prior reason to expect an inverse correlation between model and observations other than random sampling so it is reasonable to treat negative correlation as indicating no significant relationship. For this reason negative correlations are not shown in fig. 4.

Figure 4 shows anomaly correlation at each grid point, by forecast start month. It can be seen that the spatial patterns are quite similar across the direct model and all calibration methods.

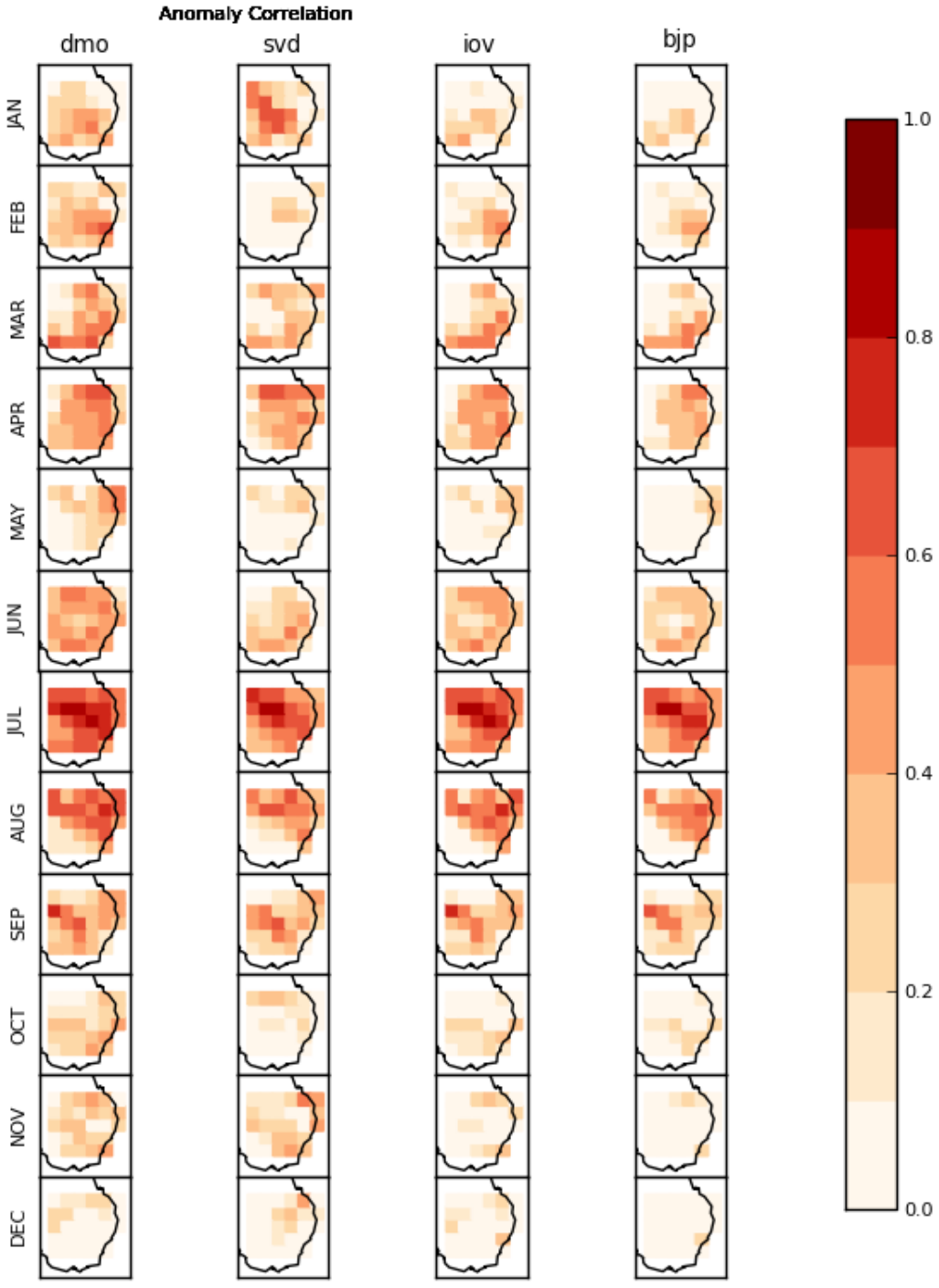

Fig. 4 Anomaly correlation for each calibration method.

Figure 5 shows the mean anomaly correlation across the region compared for the different methods. Because a correlation of approximately 0.3 is significant at the 95% level by a one tailed t-test for a time series of length 27, this is plotted as a threshold of 'useful skill', although this is still quite low and confirms the low predictability of MDB rainfall by this model. Notably, for seasonal forecasts initialised in July, the correlation of all methods remains high. This suggests none of the techniques degrades the phase relationship of the prediction where skill is high.

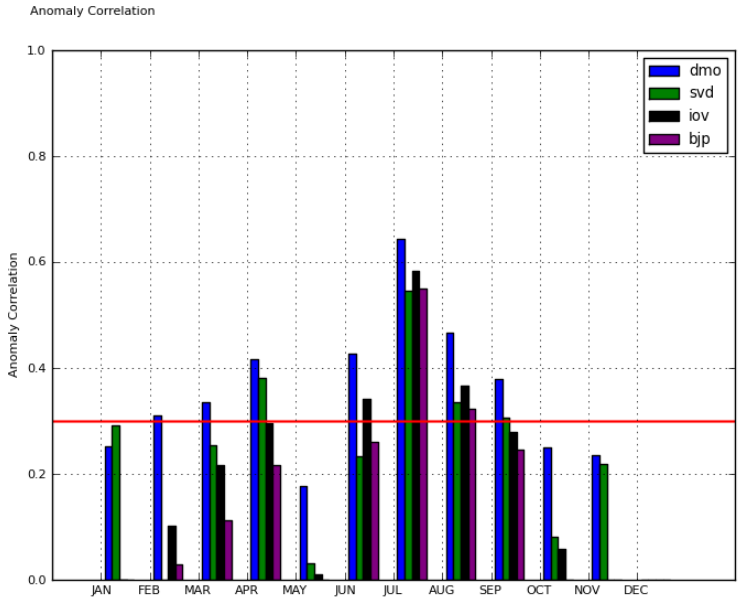

Fig. 5 Spatial mean of anomaly correlation for each calibration method. Seasonal mean rainfall.

6.2 Hit rate

Hit rate, also known as proportion correct, is a coarse grained score that measures whether events are correctly distinguished. A hit is counted for an above median event when the forecast ensemble median (or equivalently the bulk of the probability mass) is above the model's climatological median, and a miss when the forecast ensemble median is on the wrong side. Leave-one-out cross validation is used to compute the medians for the purpose of creating the probability forecast and categorising the observations as above or below median. Hit rate also is susceptible to the cross-validation artefact discussed above for low skill cases. To avoid results being affected by this artefact we treat probabilities in the range 40% to 60% as being equivocal (50%) and do not count them in the score.

The simple hit rate is open to the criticism that it scores a forecast of 51% equally to a score of 99% provided the event verifies. In other words an emphatic and an under confident forecast that lie on the correct side of 50% are scored equally as a 'hit'. However it is widely used for scoring operational forecasts and gives a clear indication of whether the forecast system is 'leaning' in the right direction. Many users are only interested in this information and so scoring forecasts as categorical is important for gauging how they will be perceived by the public. As for anomaly correlation it is insensitive to model spread, so when looking at hit rates we are examining the calibrated series for degradation and do not expect to see improvement.

Hit rate is calculated as so: if there are *a* cases where the forecast is greater than 50% and the event occurs, and *d* cases where the forecast is less than 50% and the event does not occur, with *N* total forecast cases the hit rate is given by

$$HR = \frac{\sum(a+d)}{N}. \qquad (2)$$

Figure 6 shows the hit rate by grid point. This shows similar spatial patterns to the correlation plot.

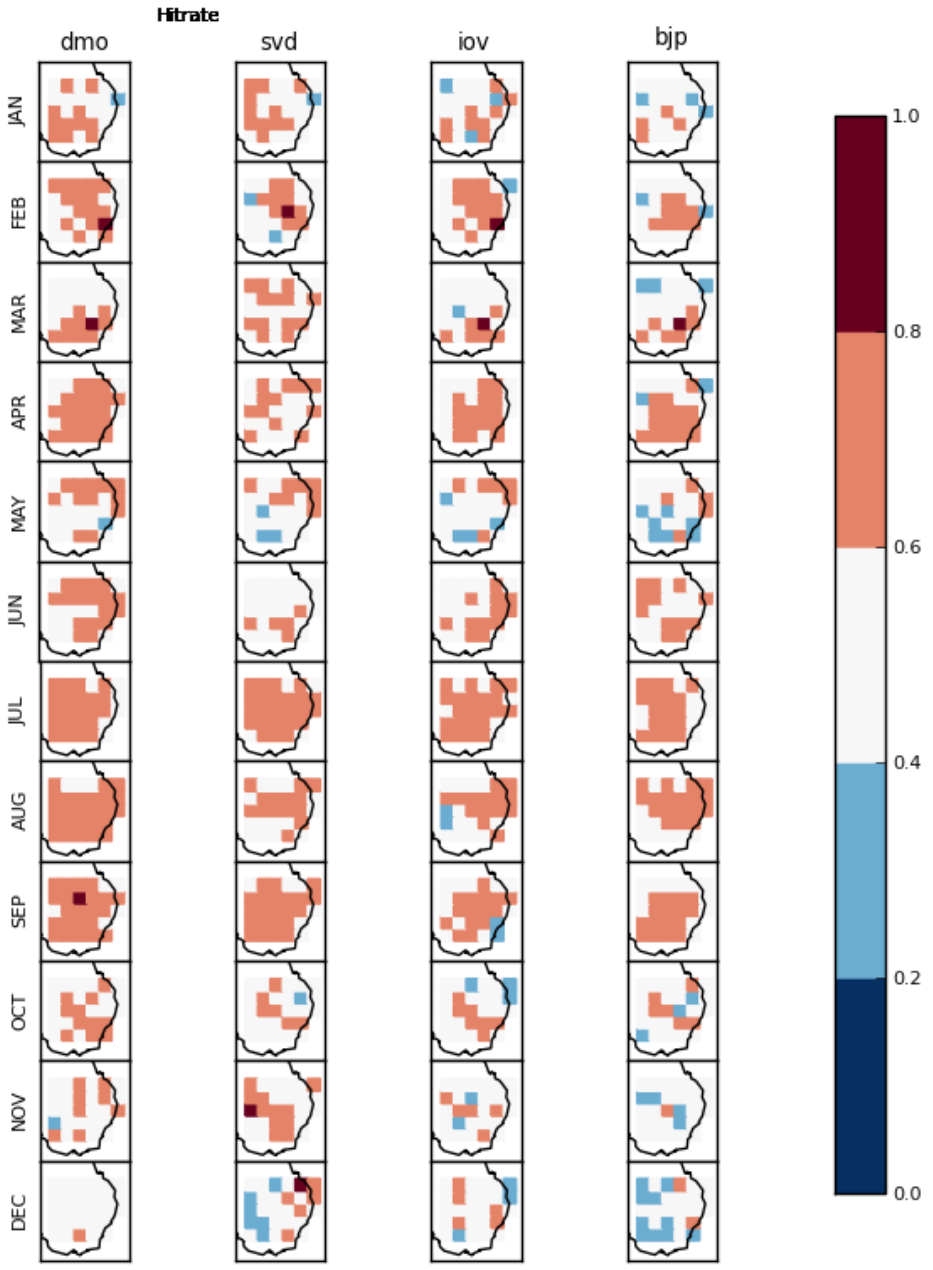

Fig. 6 Ensemble median hit rate for above median rainfall events, by grid point

Figure 7 shows hit rate averaged across the region. This figure shows that the ensemble hit rate averaged across the region is not significantly changed by calibration. The negative effect for low skill start months such as December is explained by the cross-validation artefact described above.

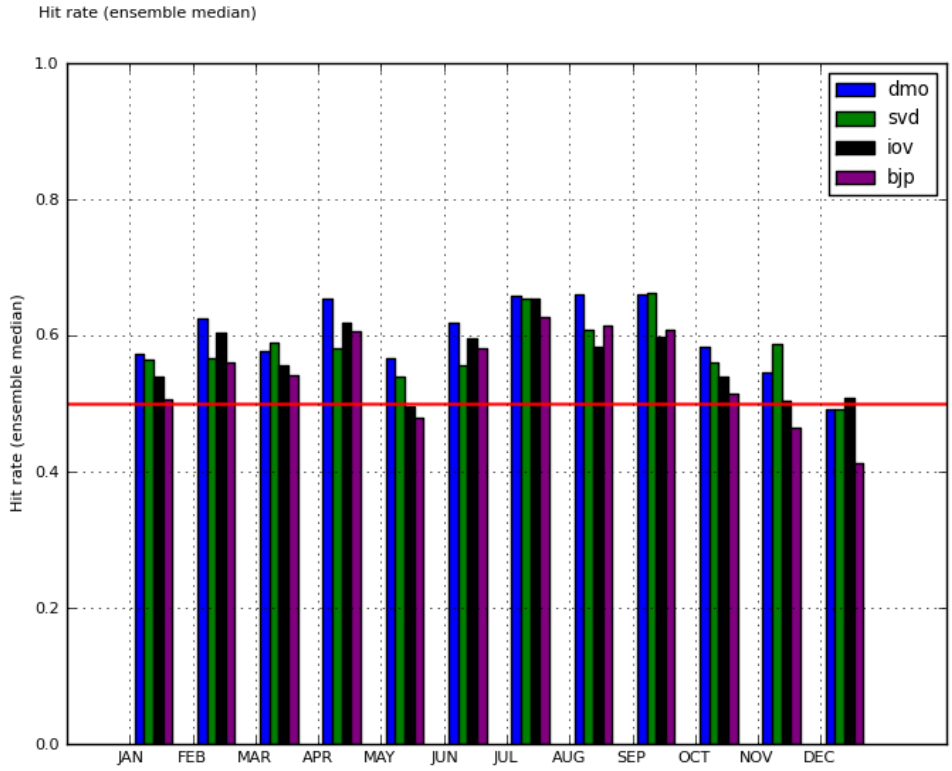

Fig. 7 Ensemble median hit rate for above median rainfall, averaged over the region.

6.3 Ensemble hit rate

To address the coarseness of the simple hit rate, an 'ensemble hit rate' can be computed. This is a count of the number of ensemble members correctly above or below the model median. Because it also makes full use of the model ensemble, this gives a smoother result than the discrete hit rate described above, and rewards emphatic forecasts that are correct.

Over all forecasts, if there are a_e ensemble members with above median rainfall and above median rainfall occurred, and d_e where both ensemble member and observations were below median with N total forecast cases and n ensemble members the ensemble hit rate is given by

$$EHR = \frac{\sum(a_e + d_e)}{Nn}. \qquad (3)$$

Figure 8 shows the ensemble hit rate by grid point. The difference between this plot and the ensemble median hit rate illustrates the increase in ensemble spread accomplished by the IOV and BJP calibrations. No calibration analysed in this study improved the ensemble hit rate. This is not unexpected for the BJP and IOV methods which bring no new information beyond the hindcast-observed relationship at each grid point. A spatial pattern correction method like SVD could be expected to improve hit rates, but we do not observe this in our study.

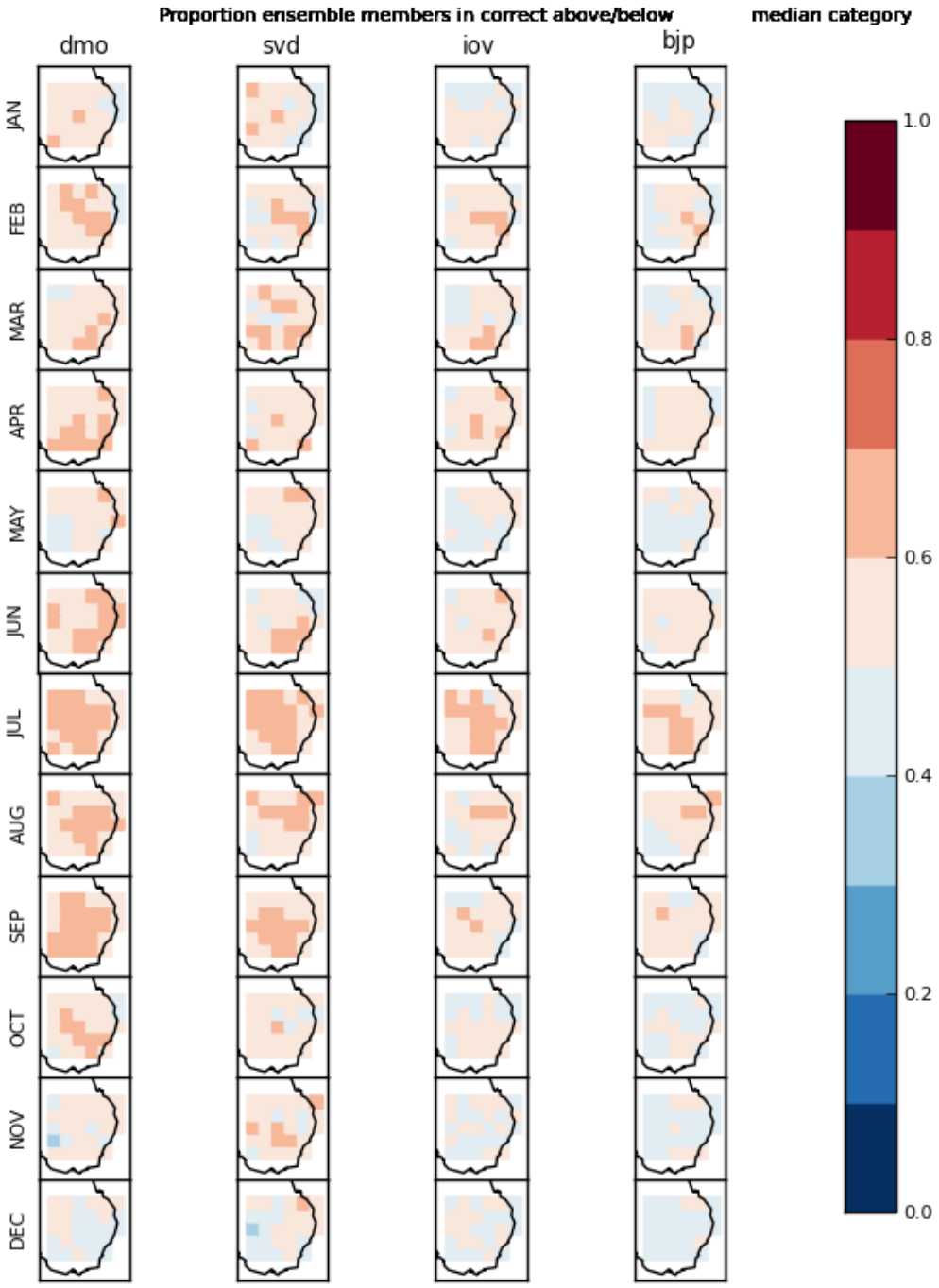

Fig. 8 Proportion correct of ensemble members by grid point.

Figure 9 shows ensemble hit rate averaged across the region. This figure shows that the ensemble hit rate averaged across the region is more or less unchanged by calibration.

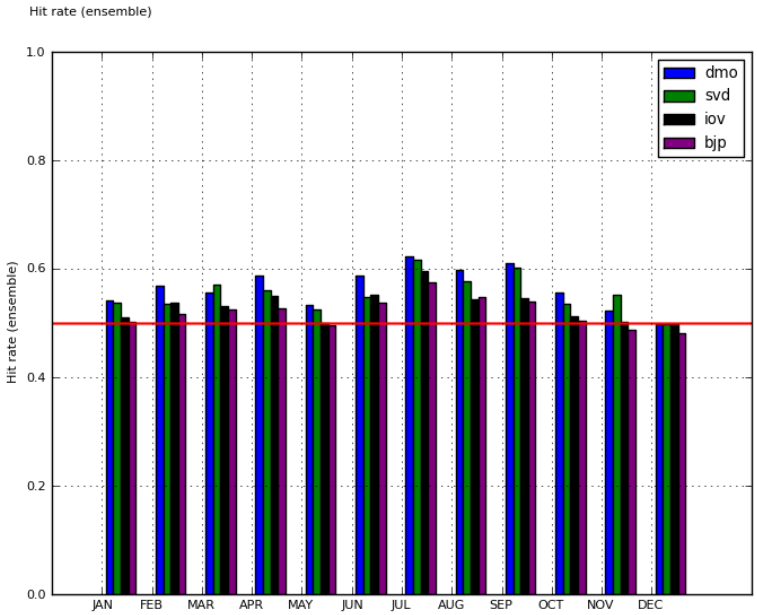

Fig. 9 Proportion correct (above/below median rainfall) ensemble members averaged over the region.

6.4 Root Mean Square Error in Probability Space

Root Mean Square Error in Probability Space (RMSEP) is similar in formulation to the widely used LEPS (Linear Error in Probability Space) score and is described in Wang et al. (2009). The RMSEP score is relatively insensitive to small numbers of large events that can dominate correlation coefficients. RMSEP will penalize forecasts that are climatologically different from observations.

The RMSEP score can be computed for the ensemble mean, median, or for the whole ensemble. We show results for the ensemble mean only. Note that because of this no information about the prediction sharpness is taken account of by this scoring. With observed cumulative climatological distribution F, N timesteps, forecast mean f and observed value o, the RMSEP score is calculated as

$$RMSEP = \sqrt{\frac{1}{N}\sum_{t}^{N}(F(f) - F(o))^2}. \quad (4)$$

A RMSEP skill score is given by the fractional improvement with respect to a climatological forecast:

$$SS_{RMSEP} = \frac{RMSEP_{clim} - RMSEP_{model}}{RMSEP_{clim}}. \quad (5)$$

Figure 10 shows the RMSEP score by grid point. There is little difference between the various calibration techniques on this metric. The BJP calibration produces the best improvement in RMSEP score, closely followed by IOV. All calibration methods seem to remove the regions of grossly poor scores apparent in January, May, June, November and December. In July the calibration methods show no improvement on this score. We speculate September season is most skilfully predicted, and the primary function of the calibration methods studied is to correct systematically bad forecasts.

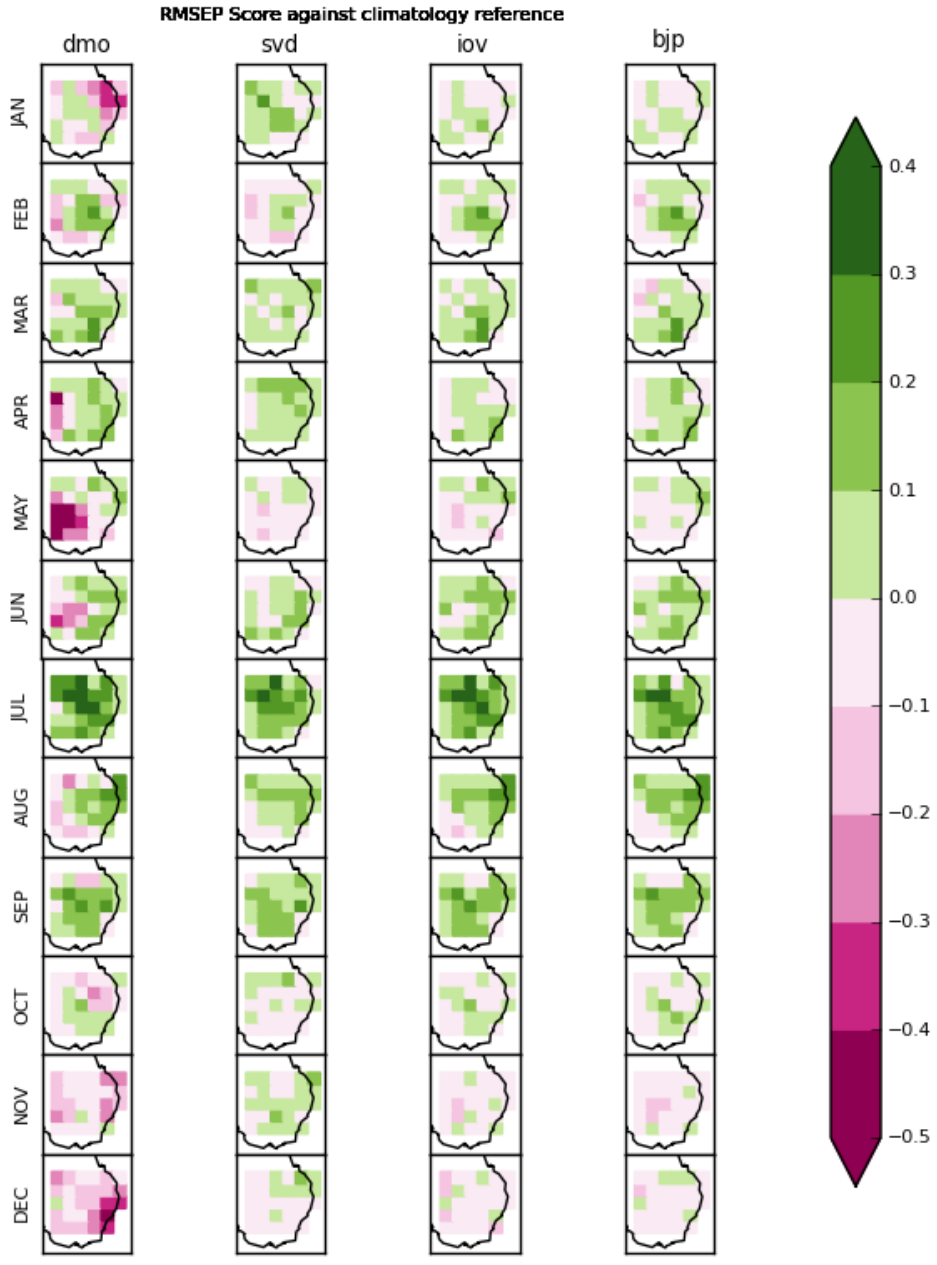

Fig. 10 Root mean square error in probability space score at each grid point.

Figure 11 shows the RMSEP score averaged across the region. Again it can be seen that the direct model scores poorly for hindcasts initialised in November, December and January. Both IOV and BJP produce a near neutral or positive RMSEP score for all months.

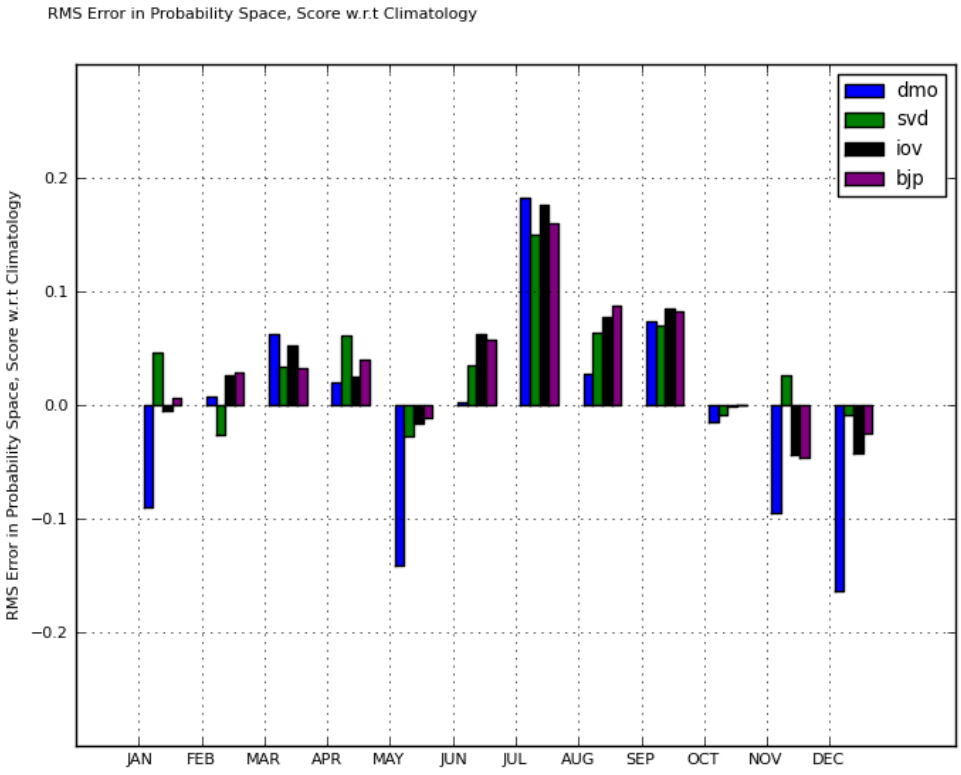

Fig. 11 Root mean square error in probability space score averaged over the region. Score is per cent improvement over a climatological forecast.

6.5 Normalised Anomaly Root Mean Square Error

Normalised anomaly Root Mean Square Error (NRMSE) is simple, intuitive and widely used, and is computed as

$$NRMSE = \frac{1}{\sigma}\sqrt{\frac{\sum\left[(f - \bar{f}_{xv}) - (o - \bar{o}_{xv})\right]^2}{N}} \qquad (6)$$

with $\bar{f}_{xv}$ indicating the cross validated mean of forecast f, observed standard deviation σ and time series length N.

The NRMSE gives the mean distance from forecast ensemble mean to the corresponding observation, indicating the magnitude of error. The error is normalised by the observed standard deviation at each gridpoint to enable comparison of errors at points of different natural variability, and to indicate where the magnitude of the error is comparable to or greater than natural variability. The coupled model output is not normalised before the error is computed. As noted above the SVD calibration includes a scaling by the cross validated observed standard deviation. Where the NRMSE is greater than one, expected error magnitude is greater than the magnitude of observed inter-annual variability.

Figure 12 shows the NRMSE by grid point. All methods have a similar NRMSE, and all calibration methods seem to remove the grosser errors visible in January and May. The spatial

NRMSE plot indicates that only forecasts starting in July have an expected RMS error less than the observed standard deviation. The spatial plots show that in general the calibrated forecasts have a more consistent error than the direct model across the region, in particular for January where the high error hotspot on the Northern coast is smoothed out by all calibrations.

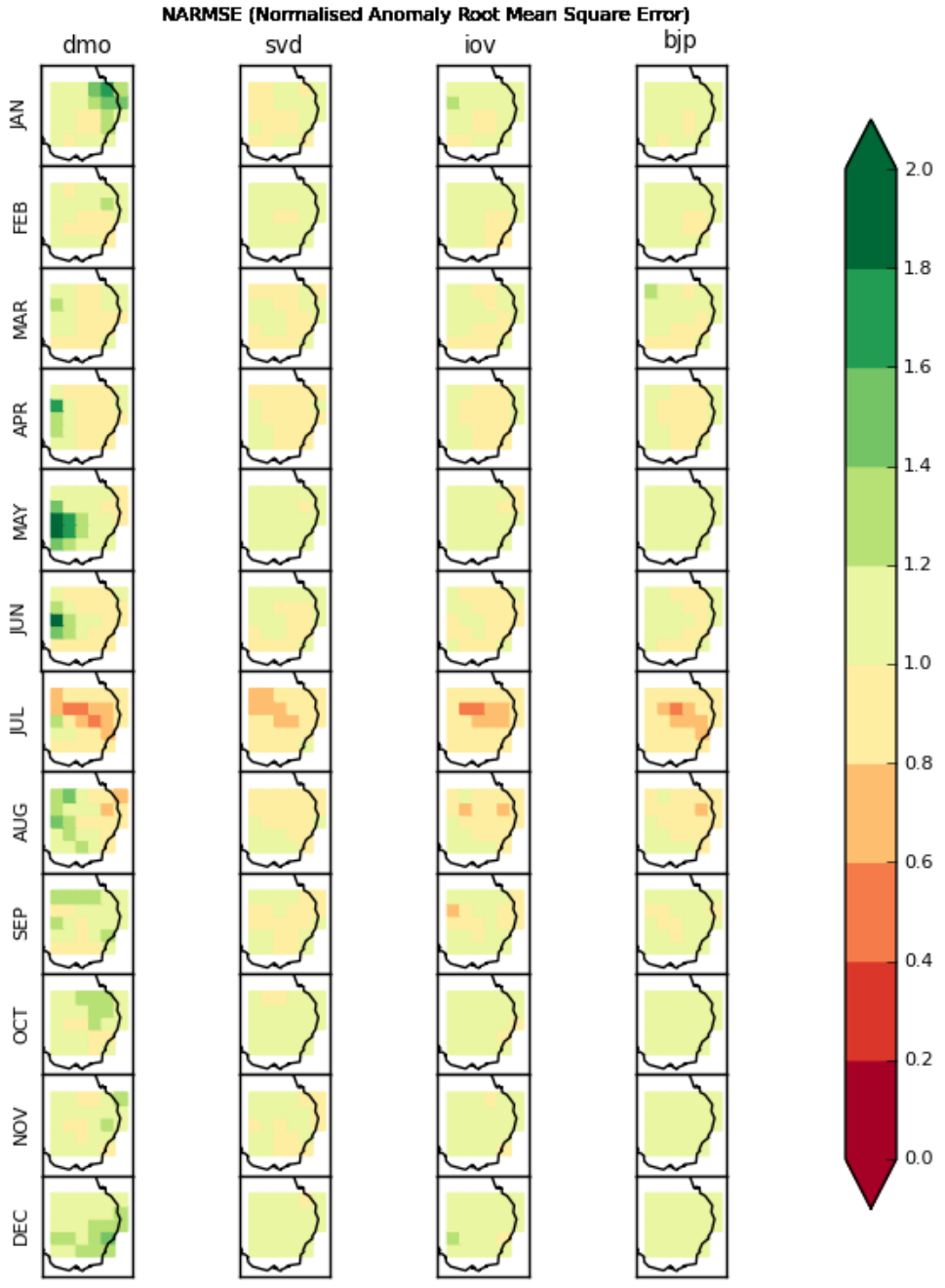

Fig. 12 Normalised RMSE for the three schemes. Values greater than 1 indicate the error is larger than the observed standard deviation (natural variability).

Figure 13 shows the NRMSE averaged across the region.

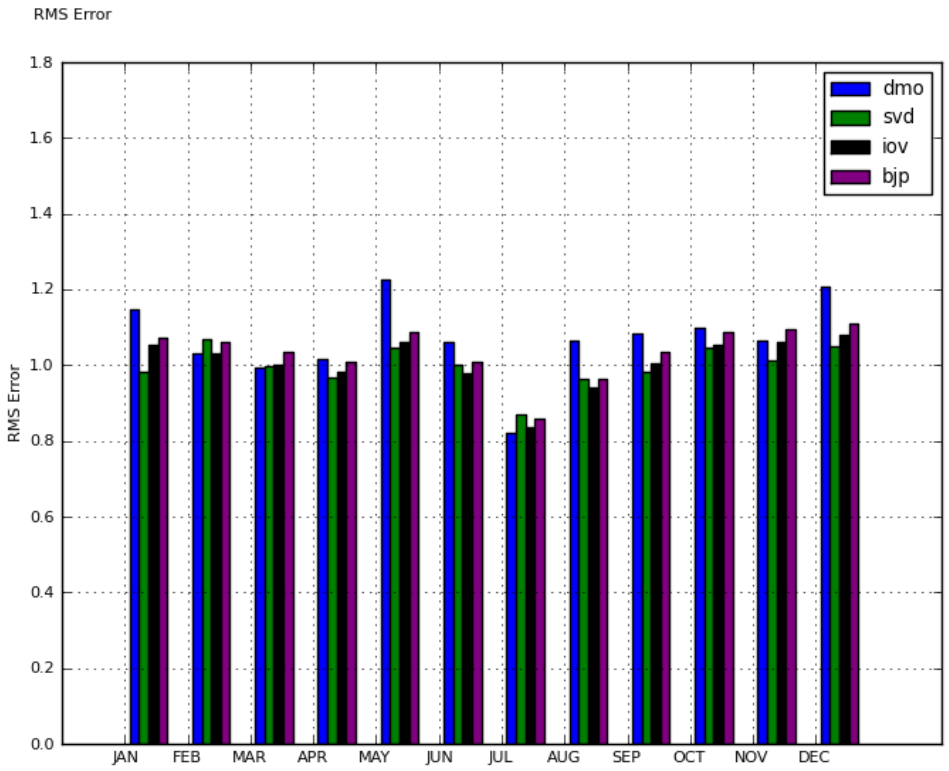

Fig. 13 Normalised root-mean-square error averaged over the MDB region

6.6 Reliability

As discussed above reliability is defined as the degree to which forecast probabilities of an event and the frequency of occurrence of the event coincide.

Figure 14 shows the reliability diagrams for the direct model and each of the calibration schemes. It can be seen that the SVD calibrated forecasts are less reliable than the direct model output. It can be seen Fig. 14 that the variance inflation improves the reliability of the ensemble forecast, but more for low probability than high probability forecasts. The BJP calibration produces the closest to perfect reliability plot of any of the analysed methods.

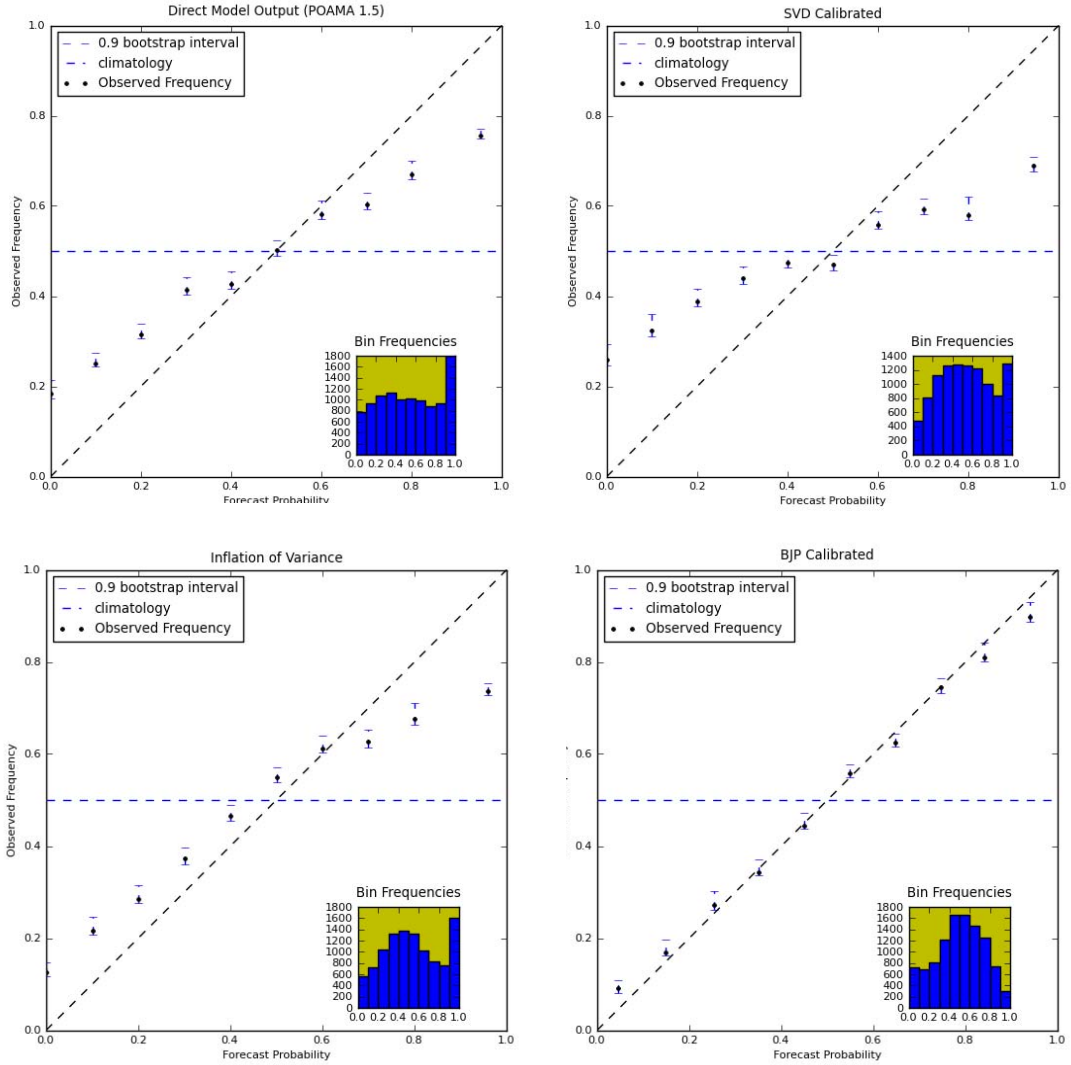

Fig. 14 Reliability over all grid points and start months Top left: direct model output. Top right: SVD calibration scheme. Bottom left: inflation of variance calibration scheme. Bottom right: BJP calibration scheme.

7. DISCUSSION

It is important to note that the three techniques used for this study are different in character and motivation: we are not clearly comparing like with like. In this section we discuss the verification results in light of each technique characteristics and elaborate on perceived strengths and weaknesses of each technique.

The SVD calibration made no improvement in reliability. For skill measures using only the ensemble mean, the SVD calibration scheme was comparable in performance to the other two methods. The SVD scheme is designed to correct errors in spatial patterns, and does not include an model error (noise) term when reconstructing the forecasts. This reduces prediction uncertainty of the third kind (due to model error). A very small number of spatial patterns are used, with most of the loading on the first two patterns. The projection onto these patterns will invariably have less variability than the original forecast. This explains why the ensemble

spread is highly truncated and there is little inter-ensemble member variability. This reduces prediction uncertainty of the first kind described in the introduction (due to internal model variability) and explains why the procedure does not improve reliability.

Another question in relation to the SVD method is whether the covariance matrix is sufficiently determined given the short time series (Lim et al. 2011).

The BJP scheme resulted in calibrated forecasts with as close to perfect reliability as could be expected.

The BJP technique performs well when the model shows no hindcast skill, generating highly reliable forecasts of climatology. The BJP method reduces the variance of the ensemble mean and hence removes the forecast signal in roughly inverse proportion to the model-observation correlation.

Grid points are calibrated as independent time series, and thus there is no way for the technique as used for this study to take account of the degree of spatial coherence of the pattern of correlation between model and observations. The modes of atmospheric variability influenced by predictable ocean circulation are large scale circulations with large scale impacts, therefore individual model grid points are not independent, especially on the seasonal timescale. Because of this, coherent regions of high spatial correlation are more likely to be related to real prediction of the circulation, while isolated points of high correlation are more likely to be random noise. The BJP describing the main important modes of rainfall variability, rather than at each grid point individually. Alternatively other methods of filtering noise in the pattern of model-reanalysis correlation might be devised.

The IOV calibration scheme improved the RMS error and RMS error in probability space to the same extent as the BJP method. The improvement to reliability of the inflation of variance scheme was positive but modest. Of the three schemes inflation of variance had the least impact on the direct model ensemble mean time series. In this sense it is a 'conservative' calibration technique that only weakly conditions on model skill. A potential problem with relying on the sample correlation is the high sampling error due to a short time series. The simplicity and ease of implementation of the IOV scheme are also strong points in its favour. As with the BJP technique the variance inflation is applied to individual grid points as if they were independent and so is susceptible to noise in the pattern of model-observation correlation. Further investigation into this question is warranted.

8. CONCLUSION

The three calibration schemes come from different families, and the hindcast period is very short. The shortness of the hindcast period, coupled with the modest level of useful skill limits the amount of useful information this type of study can provide. Variations in particular metrics for particular techniques in particular months may or may not be systematic.

The primary aim of the study was to investigate the improvement to forecast reliability of different calibration schemes. The BJP calibration scheme corrects the reliability close to perfectly and preserves the model signal in regions and periods of identifiable hindcast skill, so on this question the BJP method could be identified as a clear winner.

The inflation of variance calibration performed at individual grid points improves reliability positively and maintains the model signal, but does not improve reliability well for high rainfall forecasts.

The SVD scheme had a negative effect on reliability and so is not useful for improving this aspect of the forecasts. However we note that this family of techniques may prove useful for correcting other aspects such as spatial biases in teleconnection location, and indeed it was competitive with other methods on most metrics.

A weakness of the model output statistics approach examined here is that none of the schemes explicitly incorporate dynamical information that is relevant to prediction uncertainty such as variations in predictability with ENSO phase (which we might call 'flow dependent predictability' (Palmer and Hagedorn, 2006)). The main physical basis for seasonal prediction in the MDB is the relationship of rainfall with tropical Indo-Pacific SST variability such as ENSO. The subjective confidence of climate forecasters in dynamical model predictions is much higher during strong Pacific SST events, but work would be required to determine how to include this in an objective calibration scheme.

The short length of the available hindcasts limits how much can be learned from this exercise. It would be instructive to examine the performance of different schemes on synthetic data with known characteristics in order to better understand how the methods perform under different circumstances.

Each method analysed in this report has a particular strength. The strengths of the BJP scheme are accounting for parameter uncertainty and strongly conditioning on hindcast skill. The strength of the SVD scheme is in accounting for the large scale coherence of the main predictable modes of variability. The strength of the inflation of variance scheme is in improving the reliability with minimal conditioning on hindcast skill. A general purpose calibration scheme for generating reliable forecasts from coupled models should consider each of these elements.

9. RECOMMENDATIONS

This report recommends further investigation into general purpose schemes for calibrating dynamical model outputs into forecasts of sufficient reliability to issue publicly.

The BJP calibration method was shown in this study to generate hindcasts of rainfall in the MDB region based on POAMA 1.5 with near perfect reliability. It is recommended that an assessment the suitability of the method for the calibration of real-time forecasts from the POAMA model be carried out.

Systematic spatial biases in the POAMA coupled models are known to exist. Because of this, further research into methods of spatial pattern correction is warranted.

This report has noted the potential limitations of applying calibrations to individual grid points independently. Because of this, further research into whether spatial coherence is preserved by calibration techniques is warranted.

10. ACKNOWLEDGMENTS

This project was supported by the Water Information Research and Development Alliance (WIRADA) under project 4.2 "Improved seasonal predictions from dynamical models" and by the South Eastern Australian Climate Initiative (SEACI). Thanks to Andrew Schepen for extracting the AWAP data used for model validation. Thanks are also due to the many members

of the POAMA team and the developers and maintainers of the atmospheric and oceanic models used for the POAMA system.

REFERENCES

Alves, O., Wang, G., Zhong, A., Smith, N., Tseitkin, F., Schiller, A., Godfrey, S. and Meyers, G. (2003). POAMA bureau of meteorology operational coupled model seasonal forecast system. In *Proc. National Drought Forum*, pages pp. 49–56, Brisbane.

Barnston, A.G. and van den Dool, H.M. (1993). A degeneracy in Cross-Validated skill in regression-based forecasts. *Journal of Climate*, 6.

Broecker, J. and Smith, L.A. (2007). Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 22:651.

Coelho, C.A.S., Pezzulli, S., Balmaseda, M., Doblas-Reyes, F.J. and Stephenson, D.B. (2004). Forecast calibration and combination: A simple bayesian approach for ENSO. *Journal of Climate*, 17(7):1504–1516.

Drosdowsky, W. and Chambers, L.E. (2001). Near-Global sea surface temperature anomalies as predictors of Australian seasonal rainfall. *Journal of Climate*, 14(7):1677–1687.

Feddersen, H., Navarra, A. and Ward, M.N. (1999). Reduction of model systematic error by statistical correction for dynamical seasonal predictions. *Journal of Climate*, 12(7).

Hudson, D., Alves, O., Hendon, H. and Wang, G. (2010). The impact of atmospheric initialisation on seasonal prediction of tropical pacific SST. *Climate Dynamics*. 36:1155-1171.

Johnson, C. and Bowler, N. (2009). On the reliability and calibration of ensemble forecasts. *Monthly Weather Review*, 137(5):1717–1720.

Jolliffe, I.T. and Stephenson, D.B. (2003). *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley.

Jones, D., Wang, W. and Fawcett, R. (2009). High-quality spatial climate data-sets for australia. *Australian Meteorological and Oceanographic Journal*, 58(2):223–248.

Jones, D. and Weymouth, G. (1997). An australian monthly rainfall dataset. Bureau of Meteorology Technical Report 70.

Lim, E., 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. *Monthly Weather Review*, 139(3):958–975.

Lim, E., Hendon, H.H., Hudson, D., Wang, G. and Alves, O. (2009). Dy-namical forecast of Inter–El ni˜no variations of tropical SST and Australian spring rainfall. *Monthly Weather Review*, 137(11):3796–3810.

Murphy, A.H. and Winkler, R.L. (1987). A general framework for forecast verification. *Monthly Weather Review*, 115(7):1330–1338.

Palmer, T. and Hagedorn, R. (2006). *Predictability of weather and climate*. Cambridge University Press.

Roulston, M.S. and Smith, L.A. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(6):1653–1660.

Smith, N.R., Blomley, J.E. and Meyers, G. (1991). A univariate statistical interpolation scheme for subsurface thermal analyses in the tropical oceans. *Progress In Oceanography*, 28(3):219–256.

Spillman, C., Alves, O., Hudson, D. and Charles, A. (2009). New operational seasonal SST products for prediction of coral bleaching in the great barrier reef. *Bulletin of the Australian Meteorological and Oceanographic Society*, (22).

Stephenson, D.B., Coelho, C.A.S., Doblas-Reyes, F.J. and Balmaseda, M. (2005). Forecast assimilation: a unified framework for the combination of multi-model weather and climate predictions. *Tellus A*, 57(3):253–264.

Wang, G., Hudson, D., Alves, O., Hendon, H., Liu, G. and Tseitkin, F. (2008). SST skill assessment from the new POAMA-1.5 system. *BMRC Research Letters*, 8:1–6.

Wang, Q.J., Robertson, D.E. and Chiew, F.H.S. (2009). A bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites. *Water Resources Research*, 45:18 PP.

Yeo, I. and Johnson, R.A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954 –959.

Zhao, M. and Hendon, H.H. (2009). Representation and prediction of the indian ocean dipole in the POAMA seasonal forecast model. *Quarterly Journal of the Royal Meteorological Society*, 135(639):337–352.

The Centre for Australian Weather and Climate Research is a partnership between CSIRO and the Bureau of Meteorology.