Australian Government
Bureau of Meteorology

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

CSIRO

Assessment of international seasonal rainfall forecasts for Australia and the benefit of multi-model ensembles for improving reliability.

Sally Langford and Harry H. Hendon

CAWCR Technical Report No. 039

September 2011

www.cawcr.gov.au

Assessment of international seasonal rainfall forecasts for Australia and the benefit of multi-model ensembles for improving reliability.

Sally Langford and Harry H. Hendon

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

CAWCR Technical Report No. 039

September2011

ISSN: 1836-019X

National Library of Australia Cataloguing-in-Publicationentry

Author: Sally Langford and Harry H. Hendon

Title: Assessment of international seasonal rainfall forecasts for Australia and the benefit of multi-model ensembles for improving reliability.

ISBN: 978-1-921826-57-3(PDF/Electronic Resource)

Series: CAWCR technical report; 39

Notes: Included bibliography references and index

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

Enquiries should be addressed to:

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

S.Langford@bom.gov.au

Copyright and Disclaimer

CSIRO and the Bureau of Meteorology advise that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law, CSIRO and the Bureau of Meteorology (including each of its employees and consultants) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

Water Information Research and Development Alliance (WIRADA)
Project 4.2 “ Improved climate predictions at hydrologically-relevant time and space scales from the POAMA seasonal climate forecasts.”

The ENSEMBLES data used in this work was funded by the EU FP6 Integrated Project ENSEMBLES (Contract number 505539), whose support is gratefully acknowledged.

Contents

List of Figures

List of Tables

1. ABSTRACT

In this report we assess forecasts from Predictive Ocean Atmosphere Model for Australia (POAMA) in comparison to international dynamical coupled model forecast systems, which are archived as part of the ENSEMBLES project. We investigate how universal the lack of reliability is in dynamical forecasts of regional rainfall, in order to highlight any potential for improvement of the POAMA system. The systems assessed in this report show that overconfidence and lack of reliability for regional rainfall forecasts is a common problem. Due to the clear need for improved reliability and more accurate seasonal rainfall forecasts for hydrological applications, we have explored the benefit of combining a range of operationally available models into a multi-model ensemble, which can cancel uncorrelated error, increase spread and reduce model error. Our results indicate that there is benefit in adding POAMA version P24 to the operational models from the European Centre for Medium-Range Weather Forecasts (ECMWF), the UK Meteorological Office (UKMO) and Météo France (MF), into an equally weighted multi-model ensemble, to increase the reliability and consistency of accurate regional rainfall forecasts.

2. INTRODUCTION

Seasonal forecasts from the POAMA prediction system tend to be overconfident and only moderately reliable for predicting seasonal mean Australian rainfall (e.g. Lim et al. 2010). Although the rainfall forecasts demonstrate skill, for instance as measured by a significant correlation of the ensemble mean forecast with the observed or by a reduced mean square error relative to a reference climatological forecast, low reliability and over confidence are impediments to uptake of the forecasts for practical application. In order to assess the prevalence of this problem in other comparable dynamical coupled model forecast systems, we have investigated forecast skill and reliability for regional Australian rainfall from a range of international coupled models using hindcast data accessed from the European Union ENSEMBLES project (e.g.Hewitt and Griggs 2004, Weisheimer *et al.* 2009). Due to the clear need for improved reliability and more accurate seasonal rainfall forecasts for hydrological applications, we have also explored the benefit of combining a range of operationally available models into a multi-model ensemble, with equal weighting.

A reliable model, when considered over many forecasts, shows agreement between forecast probabilities and the frequency with which an event is observed, indicating appropriate uncertainty. Ensemble prediction systems seek to represent the uncertainty inherent in initial conditions by initiating the forecasts from a number of initial states which are perturbed from the observed state in order to represent observational uncertainty. However, simply initiating a single model with a range of perturbed initial states does not account for the full range of forecast outcomes because of model error in representing the dynamical processes. As a result, single models tend to not be reliable (Palmer *et al.* 2004) because they tend to underestimate the spread of the uncertainty, and are overconfident (Weigel *et al.* 2009). POAMA version 1.5 has previously been shown to be poorly to moderately reliable for predicting above median Australian rainfall in SON at lead time zero months even though forecast accuracy (e.g. as measured by correlation or hit rate) is typically highest in this season (Lim *et al.* 2010).

In this report we assess forecasts from POAMA version 1.5 (P15b) and 2 (P24) in comparison to forecasts from models from ECMWF, MF and UKMO that are archived as part of ENSEMBLES to determine how universal the lack of reliability is in dynamical forecasts of regional rainfall. We also address forecast accuracy in order to highlight the potential for improvement of the POAMA system. The archive of ENSEMBLES hindcasts provide an opportunity to assess the performance of coupled model forecast systems with higher vertical and horizontal resolution compared to the current POAMA versions and which might be indicative of any performance gain that will be achieved in the future when the ACCESS model is incorporated into POAMA. Because of the widespread interest in using multi-model ensemble forecasts for hydrological and other practical applications, we describe the process to access the hindcast sets from the ENSEMBLES archive at the ECMWF.

Combining different versions or independent models into a large ensemble has previously been used in operational weather prediction and is discussed in the literature as a method for increasing forecast reliability and accuracy (e.g., Weigel *et al.* 2009, 2008, Hagedorn*et al.* 2005). For example, the combination of the overconfident and emphatic forecasts taken directly from P15b with two statistical post-processed versions of these forecasts into a multi-model ensemble significantly increases the skill and reliability of the predictions (Lim *et al.* 2010). The improvement of a multi-model ensemble (MME) over a single model depends on how the independent information in each contributing model reduces forecast errors, and increases the spread in the ensemble (Weigel *et al* 2009, Lim *et al*. 2010).

The main advantage of the multi-model ensemble is that it is skilful more consistently than the single models, assuming the contributing forecast models have some level of comparable skill and the model which is most skilful varies across season and lead time, as is typical of independent models with a range of strengths and weaknesses.The multi-model ensemble provides the most consistently skilful operational system because of cancellation of uncorrelated error. It is not possible for the multi-model ensemble to perform worse than all of the individual models, as the additional information over the worst will only improve the prediction (Hagendorn *et al.* 2005). Note that deciding that a model is consistently worse than the other models would be a basis for excluding it from a multi-model ensemble, but this is a subjective assessment. A multi-model ensemble benefits from the strengths of a range of independent models, however, issues common to the models will not be resolved without further development of the parameterisation of the physical processes.

The skill of an ensemble forecast generally increases with an increase in the number of members, due to more accurate representation of the uncertainty in the initial conditions. However, it has been suggested that a multi-model ensemble with the same number of members as a single operational model will out-perform the individual model in reliability and resolution measures. For example, three or more models were shown to beat a single model with the same number of ensemble members (Hagedorn *et al.* 2005).

A number of international projects are exploring the benefits of combining the output of international centres into a multi-model ensemble forecast (Weisheimer*et al.* 2009, Palmer *et al.* 2004, Hagedorn *et al.* 2005, Doblas-Reyes *et al.* 2000). EUROpean multi-model Seasonal to Inter-annual Prediction (EUROSIP) is an operational multi-model seasonal forecast system, launched in late 2005. It combines ensemble forecasts of coupled models from ECMWF, UKMO and MF, which have been run in a consistent manner.EUROSIP currently provides

Niño plumes and spatial maps of standardized ensemble means and probabilistic forecasts to the public. Each model has 41 ensemble members for the operational forecast. Operational data is available by agreement with the contributing research centres. Currently, hindcasts from these centres are available for a smaller number of ensemble members through the ENSEMBLES project. In this report we assess POAMA in comparison to the models from these contributing centres, and explore the benefit of including the Bureau of Meteorology's forecasts into the EUROSIP project.

2.1 POAMA data

The current operational version of POAMA is P24, which combines the 10 ensemble members from the three model versions P24a, P24b and P24c into a multi-model ensemble comprising 30 members. The P15b version of the model is a 10 member ensemble (e.g., Zhao and Hendon 2009; Hudson et al. 2011). The main difference in the newer versions of POAMA is improved ocean data assimilation (Yin et al. 2010). Version P24c is identical to the version in P15b. Vesion P24a uses a slightly different version of the shallow convection scheme that results in less mean state drift and P24b uses an explicit flux adjustment to control the mean state. Hindcasts from POAMA are initialized on the first of each month for 1980-2005. Seasonal mean forecasts for MAM, JJA, SON and DJF for lead times of one and four months are assessed in this report, as this corresponds to the data available from the international coupled models. A lead time of one month corresponds to the forecast being initialised one month before the start of the verification season. For example, the lead time one month MAM forecast is initialised on the 1st of February.

2.2 ENSEMBLES data

Hindcasts from the ECMWF, Météo France and UK Met Office seasonal forecast systems are available through the ENSEMBLES project. Although there are additional models in the ENSEMBLES project, we concentrate on these three because operational predictions from systems similar to them contribute to the EUROSIP project and are thus indicative of what can be achieved currently using real-time systems. For instance The ECMWF IFS is similar to the ECSys3 seasonal forecast system that is the operational version used at ECMWF March 2007-present. Similarly, the UKMO HadGEM2 system and the MF ARPEGE/OPA system are similar to their current operational seasonal forecast systems.

The EU ENSEMBLES project, a collaboration of around 80 institutions, is focused on the benefits of ensemble prediction for seasonal to decadal forecasts. This report focuses on the multi-model and perturbed physics simulations from Research Theme 1 (RT1 – Development of the Ensembles Prediction Systems), which includes data fromthe ECMWF, MF, UKMO, the Euro-Mediterranean Centre for Climate Change (CMCC-INGV) and the Leibniz Institute of Marine Sciences at Kiel University (IFM-GEOMAR). Data is available for 1960-2005, for four start dates (Feb 1, May 1, Aug 1, Nov 1) and 7 month hindcasts (14 months for Nov 1). Table 1 outlines the components of the models considered within this project.

Table 1 Details of ENSEMBLES models with relevant references (Weisheimer et al. 2009, Rajeevan et al. 2011). The ECMWF model includes land surface modules and climatological sea-ice cover. The MF model includes the GELATO sea ice model (Salas Mélia 2002), and is coupled using version 3 of the OASIS coupler. The UKMO HadGEM2 model includes a fully interactive sea ice module. The CMCC-INGV model includes a dynamical snow-sea ice model and a land surface model.

	Atmosphere	Ocean	References
ECMWF	IFS CY31R1; T159/L62	HOPE; 0.3°-1.4°/L29	Balmaseda *et al* 2008, Stockdale *et al* 2011
MF	ARPEGE4.6; T63/L19	OPA8.2; 0.5°-2°/L31	Daget *et al* 2009, Madec *et al* 1998
UKMO	HadGEM2-A; N96/L38	HadGEM2-O; 0.33°-1°/L20	Collins *et al* 2008
CMCC-INGV	ECHAM5; T63/L19	OPA8.2; 0.5°-2°/L31	Alessandri *et al* 2010, Madec *at al* 1998, Roeckner *et al* 2003
IFM-GEOMAR	ECHAM5; T63/L31	MPI-OM1; 1.5°/L40	Keenlyside *et al* 2005, Roeckner *et al* 2003
UKMO (DePreSys)	HadAM3; 2.5°x3.75°/L38	HadOM; 1°/L40	Gordon *et al* 2000, Pope *et al* 2000, Collins *et al* 2001

Table 2 Variables in the files corresponding to monthly mean 24 hour precipitation hindcasts from the ENSEMBLES project.

Variable	Input dimensions	Description
longitude	longitude	Degrees east, size 144, equally spaced 2.5° grid
latitude	latitude	Degrees north, size 73, equally spaced 2.5° grid
reftime	time	Forecast reference time, Days since 1950-01-01 00:00:00.
leadtime	time	Time elapsed since the start of the forecast in hours.
time_bnd	time, time_bnd	Start and end of period over which the forecast(time) is valid.
realization	ensemble	Number of the simulation in the ensemble, 54 in total. In order of institutions in Table 3.
experiment_id	ensemble, string4	Experiment identifier.
source	ensemble, string60	Method of production of the data.
institution	ensemble, string15	Institution responsible for the forecast system.
sc		Height in m.
prlr	time, ensemble, latitude, longitude	Total precipitation, in m/s.

Table 3. List of the institutions and methods of production of the data for the ENSEMBLES project. Realization and Experiment ID variables as in Table 2.

Institution	Source	Realization	Exp ID
ECMWF	IFS31R1/HOPE-E, Sys 1, Met 1, ENSEMBLES	0-8	2001
IFM-GEOMAR	ECHAM5 T63L31/MPIOM GR15L40, Sys 1, Met 10, ENSEMBLES	0-8	2001
Météo-France	ARPEGEClimate4.6/OPA8.2/GELATO, Sys 0, Met 1, ENSEMBLES	0-8	2002
UK Met Office	HadGEM2, Sys 1, Met 1, ENSEMBLES	0-8	2025
CMCC-Bologna	ECHAM5/OPA8.2, Sys 1, Met 0, ENSEMBLES	0-8	2001
UK Met Office	DePreSys HadCM3+flux_cor+per_par, Sys 51, Met (10-18), ENSEMBLES	0	2502

The ENSEMBLES global hindcast atmospheric variables were accessed using Open-source Project for a Network Data Access Protocol (OPeNDAP) technology. The following link directly accesses the relevant monthly mean atmospheric fields for the stream 2 hindcasts: http://ensembles.ecmwf.int/download/ensembles/stream2/seasonal/atmospheric/monthly/. See Table 5 in Appendix A, for details of the fields available at this link.

For instance, the monthly mean 24 hour precipitation field is in the folder numbered 228. Each file name in this folder consists of an indication of the field and the start date of the hindcast. For example FC_228_mon_19970801.nc is the monthly mean 24 hour precipitation for seven months hindcast, starting from August 1, 1997. This file includes the nine ensemble members from each of the six institutions concatenated into a single dimension. The variables in a typical precipitation file are shown in Table 2. The contributing institutions and corresponding models and realisations are shown in Table 3.

ECMWF, UKMO, CMCC-INGV and MF used small perturbations of the wind stress and SST fields to generate the nine ensemble members from three different ocean analyses. The control ocean assimilation was forced by ERA-40 (Uppala *et al.* 2005) momentum, heat and mass flux data. The two other states were generated with additional wind stress perturbations based on the difference between ERA-40 and CORE (Large and Yeager 2004) analyses. SST uncertainties were represented via the addition or subtraction of the difference between Reynolds2DVAR and ReynoldsOIv2 (Reynolds *et al.* 2002). The atmosphere was initialised using ERA-40 conditions.

The IFM-GEOMAR ocean model initial conditions were determined from three coupled climate simulations with SSTs restored to observations, over 1950 to 2005. This follows the method described in Keenlyside *et al.* 2005. The nine members were generated with differing combinations of the resulting ocean and atmosphere perturbations.

The UKMO DePreSys ensemble of nine initial states was generated from perturbed surface and atmospheric parameters. The atmosphere initial conditions were taken from ERA-40 anomalies, the ocean initial conditions were determined from coupled runs which were relaxed to SST and salinity anomaly analyses.

Initialization of the models is discussed in Doblas-Reyes *et al.* 2009, Weisheimer *et al.* 2009, and further documentation is available from the ENSEMBLES RT1 web site, http://www.ecmwf.int/research/EU_projects/ENSEMBLES/exp_setup/ini_perturb/index.html.

The ENSEMBLES models have higher atmosphere model resolution than P15b and P24, but are comparable to the next version of POAMA in development. The ocean resolution is similar to the current versions of POAMA. The ECMWF atmosphere IFS model is run on a much higher resolution grid (T159 compared to T63 or equivalent for the other ENSEMBLES models and T47 for POAMA), and has finer vertical resolution (L63 compared to L17-L38 for ENSEMBLES and POAMA models).

Besides ENSEMBLES, other publicly available hindcasts from operational coupled model seasonal forecast systems includes NCEP CFSV2 T126L64 (available to download from http://cfs.ncep.noaa.gov/cfsv2.info/; Saha *et al.*, 2011). There are plans for a new multi-model project involving the ENSEMBLES and other international centres called the Climate Historical Forecast Project (CHFP) (http://www.clivar.org/organization/wgsip/chfp/chfp.php). It has commitments from about a dozen modelling centres to provide comprehensive output for at least six month hindcasts with ten ensemble members from 1979-present, starting 1st of Feb, May, Aug, Nov. This scheme is similar to ENSEMBLES stream 2. Data will be publicly available from CHFP data portal.

3 METHODOLOGY

Observed rainfall for Australia for 1980-2005 is taken from the National Climate Centre's (NCC) gridded monthly analysis (Jones and Weymouth 1997). These analyses are on a 0.25x0.25 degree longitude-latitude grid in the range 44.5ºS-10ºS, 112ºE-156.25ºE.

Seasonal mean rainfall forecasts from the different models and the corresponding observations were interpolated to the POAMA 2.5ºgrid over Australia before any analysis, using bilinear interpolation. The land and ocean mask files were also interpolated to the POAMA grid.

Deterministic forecasts are calculated from the mean of all ensemble member anomalies, as compared to the individual model climatology. When assessing the forecast skill, the computation of the model climatology, means and medians are all leave-one-out cross validated.

In this report, probabilistic forecasts of observing above median are assessed. The forecasts are calculated for each grid point, year, season and lead time from the fraction of ensemble members that predict rainfall greater than the correspondingleave-one-out cross-validated median value. In assessing the forecasts, the observations were compared to their own leave-one-out cross-validated climatological median value. This removes any biases in the mean between the models and the observations.

Due to an even number of ensemble members in the POAMA hindcasts (10 in each version of the model), there is often the situation where half the ensemble members indicate a wetter than median forecast, and half the ensemble members indicate a dryer than median forecast. This corresponds to exactly a 50 per cent probability of above median rainfall, and a 50 per cent probability of below median rainfall. As the observations only fall into a single category, the choice on how to 'award' this two category forecast can affect the skill measure in comparison to other forecast systems. If a single hit is given for an equal probability forecast, no matter which category is observed, the skill is biased higher than if a hit is not given for this situation. As the ENSEMBLES models have an odd number of ensemble members, an equal probability forecast never occurs.

Therefore, to avoid 50 per cent forecasts, a subset of nine ensemble members was used to create the POAMA forecasts. As the ensemble members are created from a set of perturbed initial conditions, we must ensure the forecast is not biased by the selection of the subset of members. To properly sample the range of initial conditions with only nine ensemble members, the probabilistic forecasts resulting from the random choice of nine members without replacement were averaged over 100 Monte Carlo simulations. For deterministic forecasts, the ensemble mean of the randomly selected nine members was averaged over the 100 Monte Carlo runs.

3.1 Multi-model ensembles

A multi-model ensemble combines the ensemble members from a number of independent models into a single prediction. For a deterministic forecast, the ensemble-member anomalies are calculated from the relevant model climatology, and then all model ensemble members are averaged to calculate the multi-model ensemble mean. The weighted ensemble mean of a multi-model ensemble with M models is therefore;

$$\overline{x}_{MME} = \frac{1}{M}\sum_{k} w_k \left(\frac{1}{N_k} \sum_{i} x_k^i \right) \qquad (1)$$

where w_k is the weight of the kth model which has N_k ensemble members, and the anomalies x_k^i are calculated using the model climatology of the individual model.

For a probabilistic forecast from a multi-model ensemble, the probabilistic forecasts from each individual model are averaged, with weighting if necessary;

$$p_{MME} = \frac{1}{M}\sum_{k} w_k p_k \qquad (2)$$

where p_k is the probability of the event occurring from the kth model forecast (Johnson and Swinbank 2008). In this report, only a simple combination is explored, where all models have equal weight (w_k=1 for all k). Averaging the probabilistic forecasts of each contributing model, as opposed to pooling the ensemble member anomalies to then create the forecast, accounts for any bias in the ensemble mean or difference in spread between the individual models.

The new operational version of POAMA consists of P24a, P24b and P24c combined into a single version of the model. The contributing versions are not independent. Hence, the multi-model ensemble for P24 benefits from an increase in ensemble members, but does not fully benefit from a large range in possible parameterisations of the physical processes.

4 HINDCAST ASSESSMENT

4.1 Accuracy

For a two-category forecast of above/below median rainfall, the contingency table (Table 4) can be used to determine the accuracy of the forecasts.A forecast where the probability of above median rainfall is greater than 50 per cent is categorised as a ‘yes’, and where it is less than 50 per cent is categorised as a ‘no’. An observation of above median rainfall is categorised as a ‘yes’ and an observation of below median rainfall is categorised as a ‘no’. As discussed in Section 3, there are no forecasts where the probability is exactly equal to 50 per cent.

Table 4 Contingency table for diatomic forecast, e.g. above/below median rainfall.

		Observed	
		Yes	No
Forecast	Yes	Hit	False Alarm
	No	Miss	Correct Negative

The accuracy score is equal to the total number of hits and correct negatives, divided by the total number of forecasts;

$$\text{Accuracy Score} = \frac{\text{Hits + Correct Negatives}}{\text{Hits + Misses + False Alarms + Correct Negatives}} \text{ x } 100\ \% \tag{3}$$

Figure 1 shows the accuracy scores of the individual models for a lead time of one month. The accuracy at each grid point is calculated from 26 years of cross-validated forecasts. An accuracy score greater than 50 per cent is considered a skilful forecast, and is represented by the green and blue grid points.

The models are most accurate for predicting regional rainfall in austral autumn (MAM) and spring (SON). In MAM, the models are typically most accurate in the north west of the continent, except P24, which is accurate in the south east. The ECMWF model has the highest accuracy in winter (JJA), although P24 is also reasonably skilful in this season. In SON, all models are accurate in the centre and east of the continent. Austral summer (DJF) is typically the least skilful season for all modelsexcept the UKMO model, which has high accuracy in Western Australia.

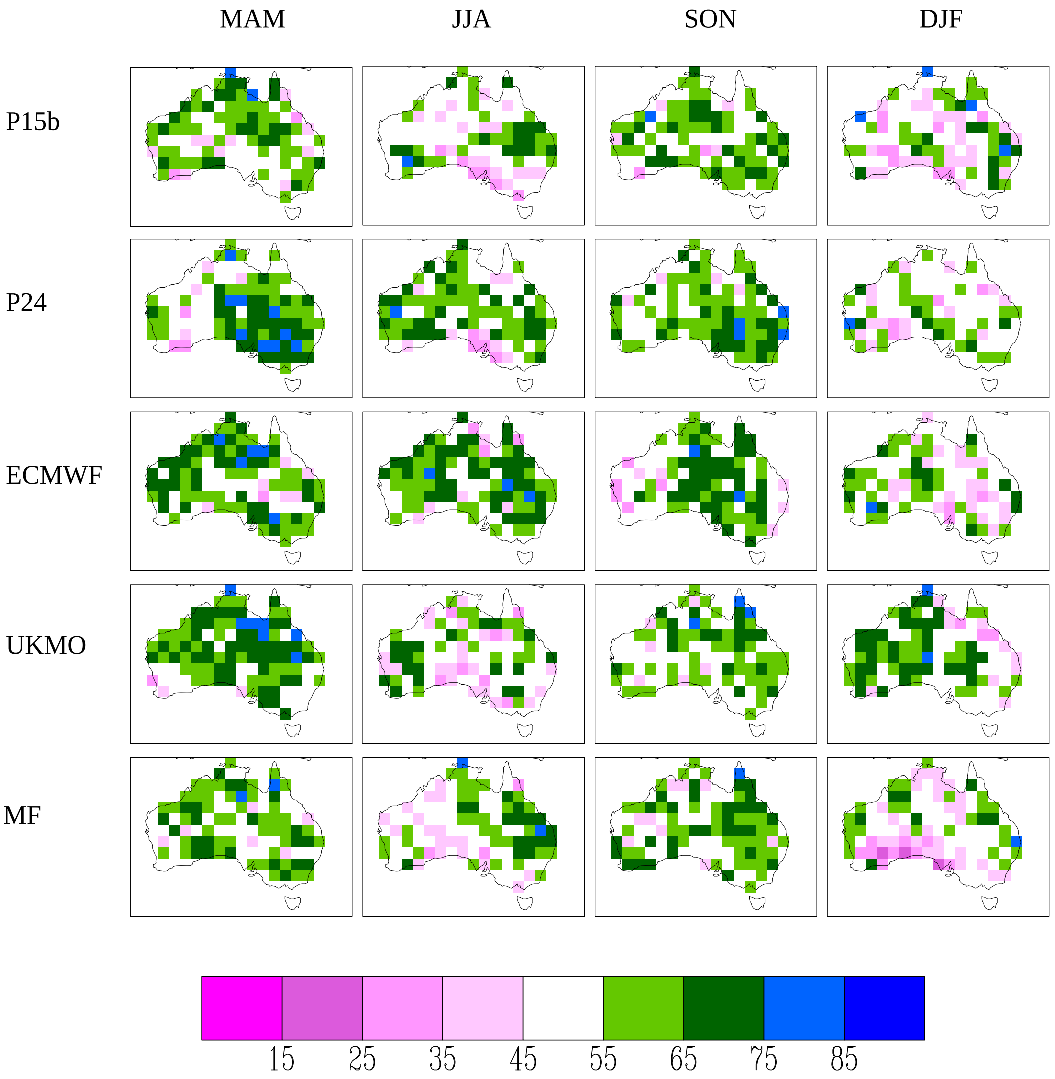

Fig.1 Accuracy score for P15b, P24, and ECMWF, UKMO and MF models, for a lead time of one month. Probabilistic forecasts for P15b, and for P24a, P24b and P24c which contribute to P24, were generated from the average of forecasts based on nine randomly selected ensemble members, as described in Section 3. An accuracy score greater than 50 per cent (indicated by green or blue grid points) is considered skilful.

SON is the most consistently skilful season for all models. This stems from the fact that El Niño-Southern Oscillation (ENSO), which is the most predictable mode of variability in the climate system, has the strongest impact on rainfall across Australia in this season (McBride and Nicholls 1983). Due to this prominence of ENSO's impact in spring, forecast skill at a lead time of four months (Fig. 2) remains high while it decreases in all other seasons.

Interestingly, the models that showed high overall accuracy in particular seasons at the short lead time are no longer the most overall accurate models at longer lead time. This suggests that the high skill at short lead timeis partly due to accurate atmosphere/ocean initial conditions, like that of the ECMWF model in JJA. While the MF model shows only moderate accuracy at the short lead time, it does not show the large decrease in accuracy with lead time seen in the other models.

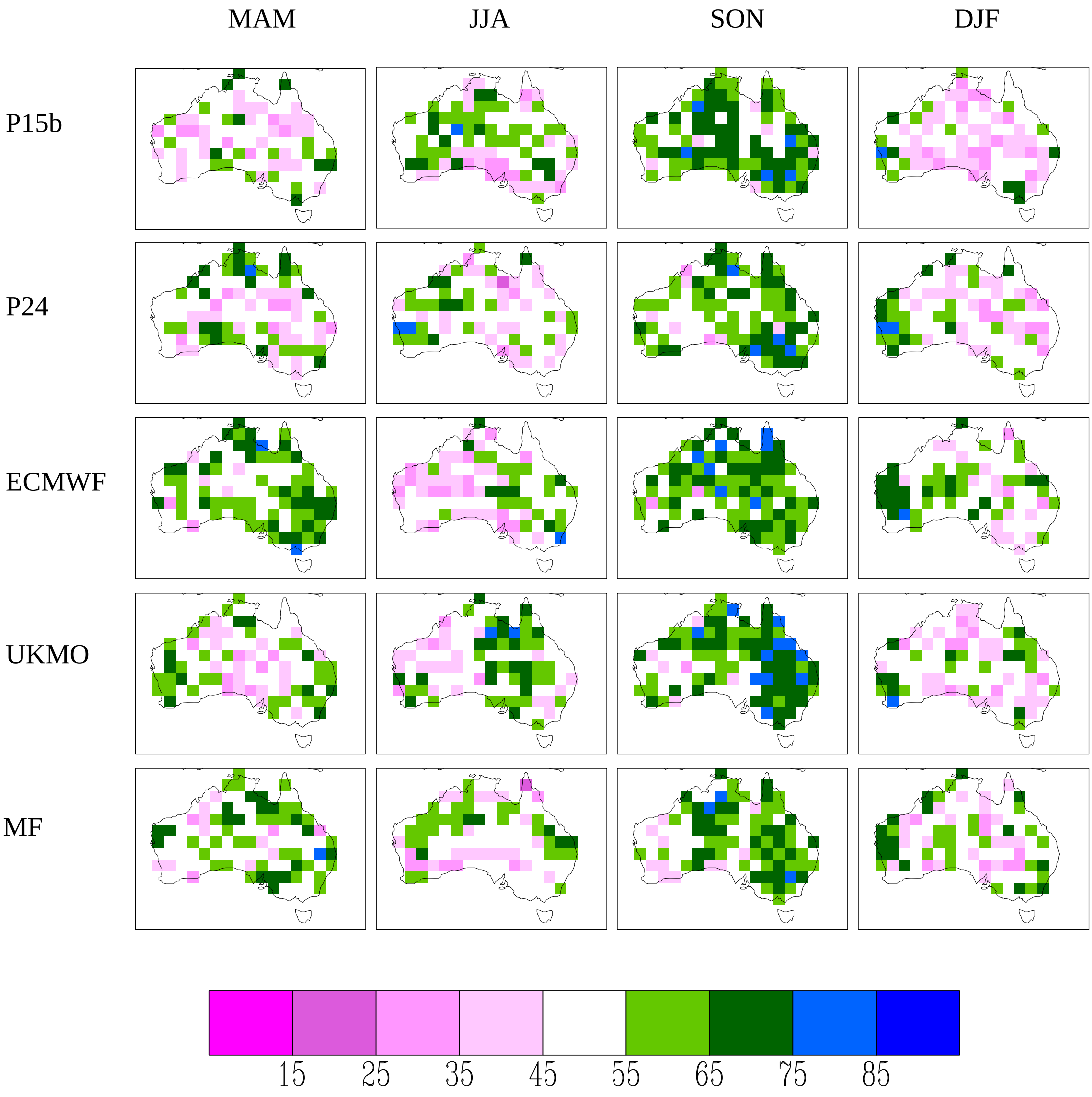

Fig.2 As for Fig. 1, for a lead time of four months.

The other ENSEMBLES models not shown here have similar patterns of common areas of accuracy. For example, they show high accuracy in the centre and east of Australia in SON, and a lack of accuracy in DJF. The INGV model has the lowest overall accuracy of the ENSEMBLES models.

As another measure of forecast accuracy, the correlation of the ensemble mean rainfall prediction from the individual models with the observations is shown in Fig. 3. Although correlation is a deterministic assessment of forecast skill, it shows similar patterns to the accuracy scores for probabilistic forecasts seen in Fig. 1. There is a positive correlation to the observations (greater than 0.6 in large areas) of the MAM and SON rainfall forecasts. There is strong positive correlation with the observations (greater than 0.7) for P24 in MAM, ECMWF in JJA, and UKMO in MAM and DJF. Strong negative correlation (less than -0.7) is seen in the south of the continent in JJA and DJF. This corresponds to regions of low accuracy (less than 25 per cent correct) in Fig. 1.

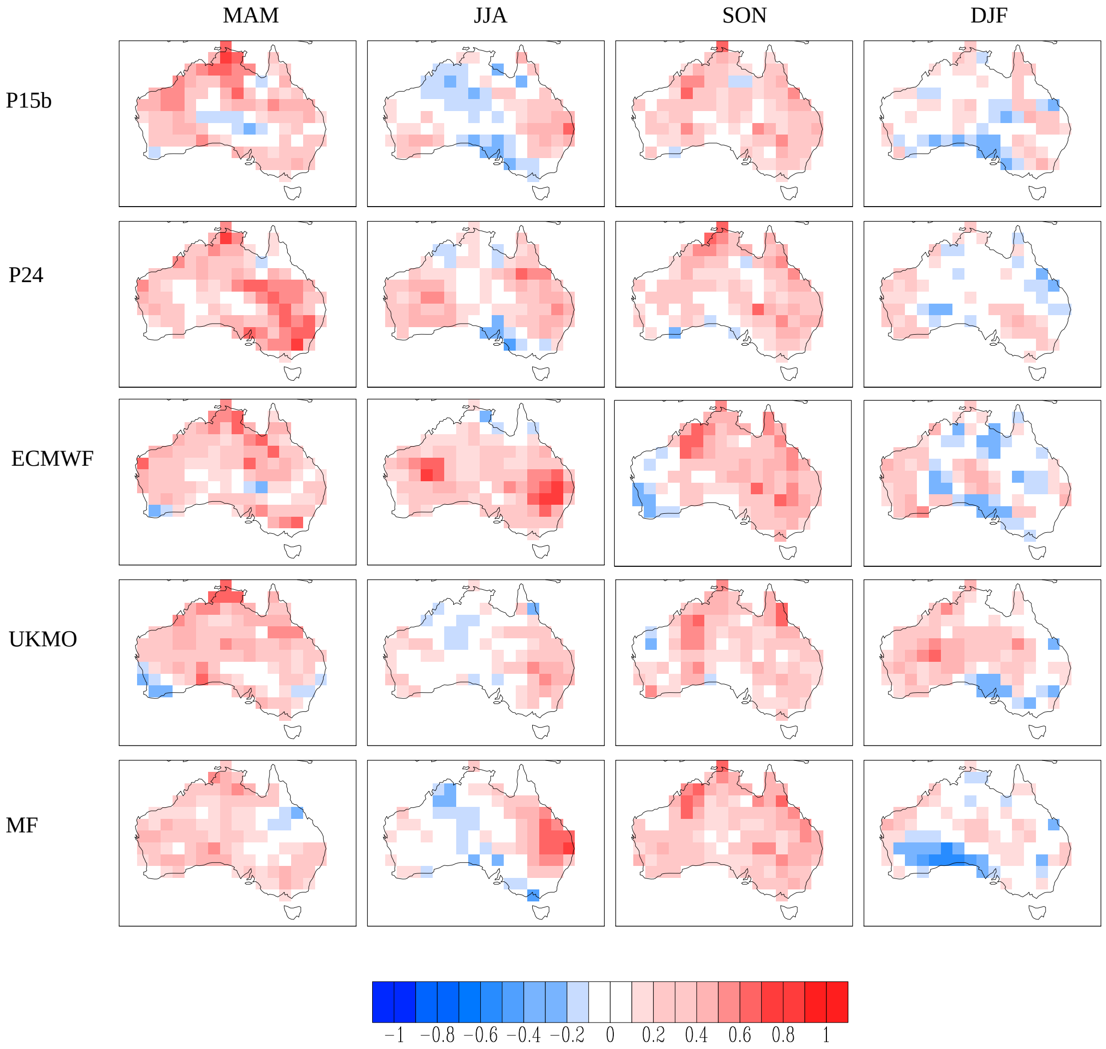

Fig.3 Correlation of ensemble mean with observationsfor the indivdiual models, for a lead time of one month.

The correlation with observations decreases with lead time for all seasons and models (not shown). The smallest decrease is seen in SON. The regions of high positive correlation in MAM and JJA in the ECMWF, UKMO and POAMA models become negative or uncorrelated with the observations at the longer lead time.

4.2 Reliability, Resolution, Sharpness

In order for a dynamical forecast to be valuable and practically applicable, it must be reliable as well as accurate. A reliable forecast predicts an event with a probability that corresponds to the frequency with which the event is observed, when considered over many forecasts. An accurate probabilistic forecast has resolution and sharpness, as well as reliability.

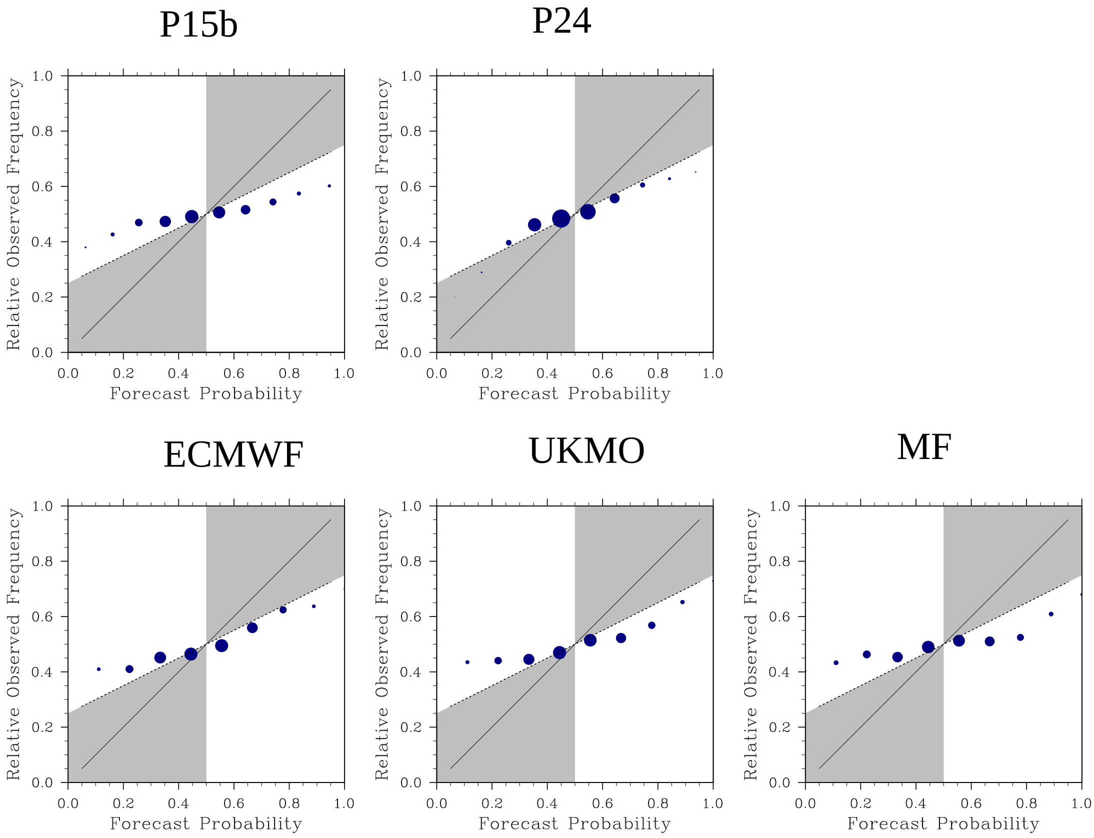

Fig.4 Attributes diagrams for the individual models, for all four seasons and two lead times combined. The y-axis is the relative observed frequency, the x-axis the forecast probability (average of forecasts within the probability bin). The solid line shows perfect reliability. The dashed line is the no-skill line, which borders the shaded area indicating skilful forecasts.

The reliability of forecasts can be represented on a reliability diagram or attributes diagram, where the relative observed frequency of an event is plotted against the forecast probability, which is divided into a number of bins (e.g., Wilks 1995). In this report, ten probability bins are equally spaced between 0 and 100 per cent. A reliable forecast will lie along the diagonal 1:1 line, such that the event is observed to occur on a fraction of occasions equal to the probability with which it was forecast. A reliability or attributes diagram requires pooling forecasts across seasons, lead times or locations in order to increase the sample size. If the forecasts from grid points are pooled into a reliability diagram, a reliable forecast will then have the event occurring at a certain fraction of locations, not just on a certain proportion of occasions.

Figure 4 shows the attributes diagrams for the individual models pooled over all four seasons and two lead times for above median rainfall forecasts using all land points over Australia. The size of the data points correspond to the fraction of forecasts in that probability interval. The dashed line indicates the no-skill line, bordering the shaded area which represents the region where a forecast will contribute positive skill to the Brier Skill Score when compared to climatology (see Equation (4) below). Note that the attributes diagrams in this report show data points corresponding to the average of the forecasts in each bin rather than the central value (Brocker and Smith 2007) so that a reliable forecast will accurately lie on the solid 1:1 line.

An attributes diagram is a reliability diagram which includes a representation of resolution and sharpness, indicated by the size of the data point in each frequencybin being proportional tothe number of forecasts in that bin. Sharpness is the tendency of the forecast to predict extreme values away from climatology. For a probabilistic forecast, this is indicated by large data points near 0 and 100 per cent. Resolution is the ability of the forecast to discriminate between events and non-events, such that a different forecast results in a different distribution of outcomes. If the data points lie along the horizontal at a relative observed fraction of 0.5, the forecast is said to have zero resolution. This means that climatology is observed, no matter what was forecast. If the data points are grouped around a forecast of 50 per cent probability, the forecast cannot discriminate from a climatological forecast, and is said to lack sharpness and resolution, although it may be perfectly reliable. Sharpness and resolution are properties of the forecast only, and an incorrect forecast may have these attributes.

Forecasts from all of the models considered here are overconfident and lack reliability and resolution, as they have a shallower gradient than the perfectly reliable (diagonal solid) line. POAMA version P24 is the most reliable individual model, and is much more reliable than the earlier version P15b. ECMWF also shows moderate reliability, with some forecasts lying on or near the no skill line. All of the models show reasonable resolution, although in individual low skill seasons such as DJF or at the longer lead time they are likely to have forecast probabilities which do not depend on the number of eventsobserved. Combining all the seasons and lead times together as we have done in Fig. 4 allows the consistency of the models to be more easily assessed, and gives a better indication of the value of the predictions because reliable forecasts are required at all times and locations. If the reliability error and resolution are assessed separately (see below), it can be seen how the models perform by season and lead time.

A measure of the reliability and resolution of a forecast can be determined from the Brier Score (BS), which measures the magnitude of the probability forecast errors. The BS can be decomposed into three components (e.g. Wilks 2006);

$$BS = \frac{1}{N}\sum_{t=1}^{N}\left(f_t - o_t\right)^2 = \frac{1}{N}\sum_{k=1}^{K} n_k\left(f_k - \bar{o}_k\right)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k\left(\bar{o}_k - \bar{o}\right)^2 + \bar{o}\left(1-\bar{o}\right) \qquad (4)$$

where N is the total number of forecasts, K is the number of probability categories, n_k is the number of forecasts in a probability category, f is the forecast probability, o the outcome of an event (equals one for an occurrence and zero for a non occurrence), $\bar{o}_k$ is the observed frequency, and ō is the observed climatology. A perfect forecast has a Brier score of zero.

Forecast skill of the model can be compared to a climatological reference forecast using the BS score as the skill measure in the Brier Skill Score (BSS), which is a measure of percentage improvement of a forecast over a reference forecast. We note, however, that the BSS is negatively biased for small ensembles, as discussed in Muller et al (2005) and Weigel et al (2007a). A bias correction can be used with the Brier Skill Score to account for the small ensemble size. For a multi-model ensemble, the bias correction is based on the weighting of the individual models (Weigel et al 2007b). However, in order to avoid this problem of negative bias with the BSS when assessing different models with differing ensemble sizes, we will not consider the BSS. Instead, we will assess the components of the unbiased Brier Score, shown in Equation 4.

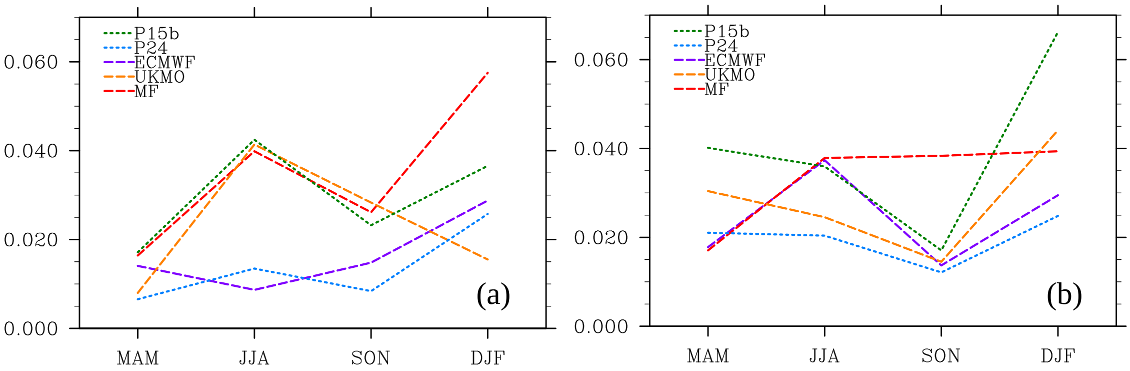

Fig.5(a) Mean reliability error for the individual models as a function of season,for a lead time of one month. **(b)** For a lead time of four months.A low mean reliability error value indicates a more skilful forecast.

The reliability error is the first term of the Brier Score (Equation 4);

$$REL = \frac{1}{N}\sum_{k=1}^{K} n_k \left(f_k - \overline{o}_k \right)^2 \qquad (5)$$

A perfect reliability error score of zero indicates a 1:1 correspondence between the forecast probability and the relative observed frequency.

The resolution is the second term of the Brier Score (Equation 4) and determines the ability of the forecast to differentiate from a climatological observation;

$$RES = \frac{1}{N}\sum_{k=1}^{K} n_k \left(\overline{o}_k - \overline{o} \right)^2 \qquad (6)$$

A forecast has good resolution if the outcome changes with a differing forecast probability. Although the forecast probability is not directly input into the resolution score, it is inherent in the relative observed frequency. Sharpness is generally the departure of the forecasts from climatology (as discussed above), but there is no mathematical formulation.

Figure 5 shows the seasonality of the reliability error of rainfall forecasts for Australian land points at lead time one and four months. MAM and SON are the most reliable seasons for all models. Winter (JJA) is unusual in that two models have relatively low reliability error but three models have relatively high reliability error. There is a lack of reliability in all models in DJF. The seasons that the individual models are more accurate are also the seasons where the models are more reliable.

POAMA version P24 (blue, dotted line) is seen to be the most reliable model in MAM and SON as it has the lowest reliability error. The ECMWF model (purple, dashed line) is the most reliable ENSEMBLES model, and slightly more reliable than P24 in JJA. In MAM and DJF, the UKMO model (orange, dashed line) shows high reliability at the short lead time, butthis decreases at the longer lead time. The MF model (red, dashed line) and POAMA version P15b (green, dotted line) show low reliability in all seasons and lead times, although the MF model does not decrease in reliability at the longer lead time by as much as the other models. At a lead time of four months, the SON forecasts remain reliable or become more reliable.

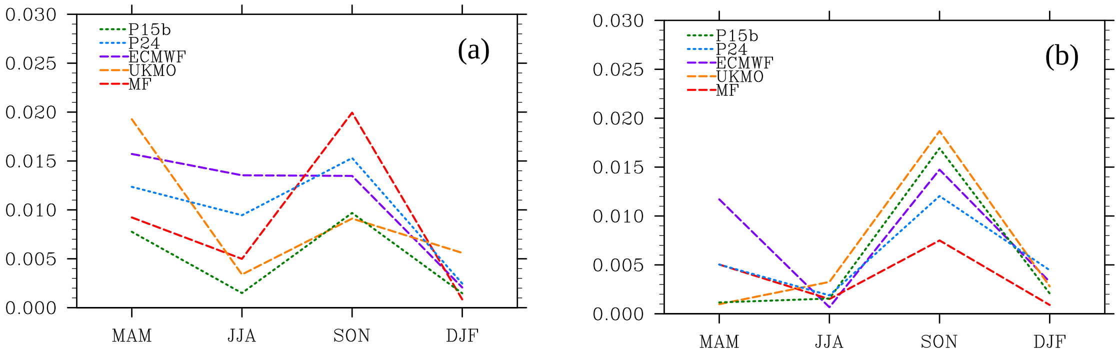

Fig.6(a) Mean resolution for the individual models as a function of season,fora lead time of one month.**(b)** For a lead time of four months. A high resolution indicates a more skilful forecast.

Figure 6 shows the seasonality of the resolution of the models at lead time one and four months. The seasonality of the resolution of the models is similar to the seasonality of the reliability error. A noticeable difference is the MF model, which has the highest resolution in SON at the shorter lead time.

POAMA version P24 does not have the highest resolution of the models, despite its high accuracy and reliability. This lack of resolution can be seen in the attributes diagrams in Fig. 4, as the larger data points closer to the 50 per cent forecast probability indicate a larger number of close to climatological forecasts than emphatic forecasts.

In summary, POAMA version P24 shows consistent improvements for forecast accuracy, reliability and resolution over P15b. Compared to the operational models that have contributed forecasts to the ENSEMBLES projects, overconfidence and lack of reliability for their regional rainfall forecast is a common problem. P24 is the most consistently reliable of the models considered here, although each of the European models shows varied strengths for particular seasons or lead times. In particular, the ECMWF model is highly accurate and reliable in JJA at the shorter lead time. The UKMO model stands out as highly accurate and reliable in MAM and DJF at the shorter lead time.

5 MULTI-MODEL ENSEMBLES

In order to assess the benefit of combining these strengths and weaknesses of the individual models, the ENSEMBLES hindcasts from ECMWF, UKMO and MF were combined into a 27 member multi-model ensemble (referred to as MME(a)). These three models were selected as they are the current EUROSIP partners, and operational versions are available to partner institutions.

These three models were also combined with the 27 member version of POAMA, P24, into a multi-model ensemble with a total of 54 members (referred to as MME(b)). The increase in accuracy and reliability can therefore be assessed compared to the individual versions of these contributing models.

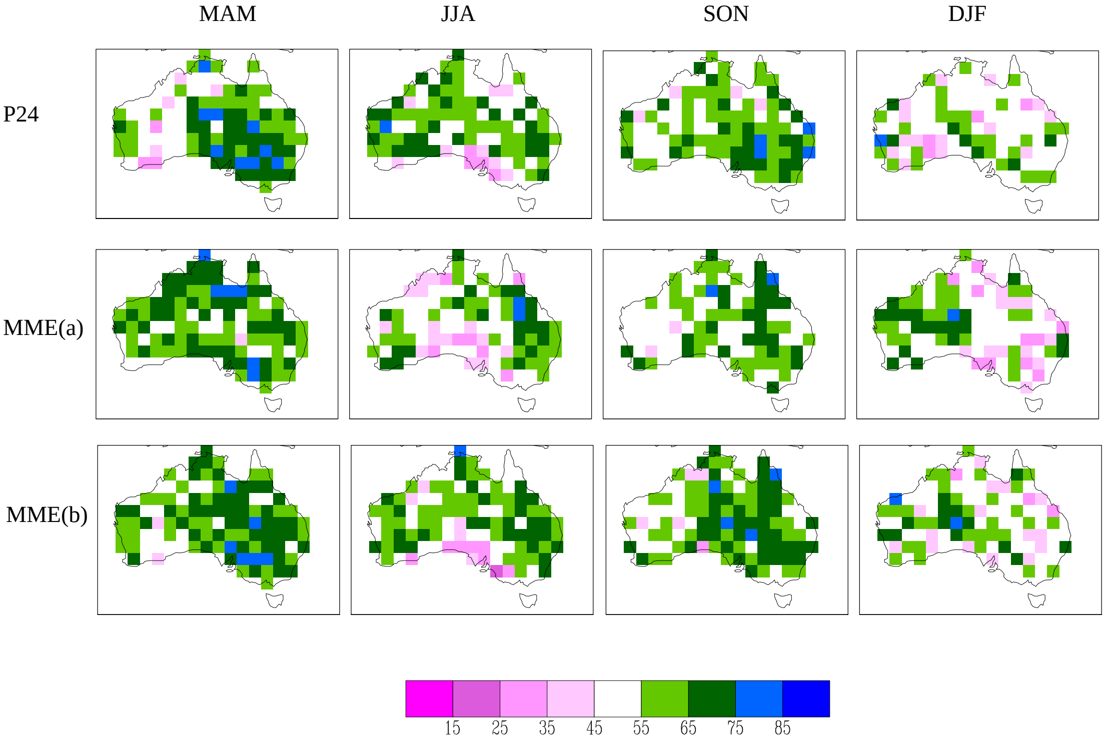

Fig.7 Accuracy score for P24 (as before) and two multi-model ensembles, for a lead time of one month. MME(a) corresponds to the multi-model ensemble of ECMWF, UKMO and MF models, with nine members each. MME(b) corresponds to the multi-model ensemble of ECMWF, UKMO and MF models, nine members each, plus the P24 model, which was created from nine members of each of P24a, P24b and P24c.

5.1 Accuracy

Figures 7 and 8 show the accuracy scores for P24 and the multi-model ensembles at a lead time of one month and four months respectively. These figures are equivalent to Figs. 1 and 2 for the individual contributing models. The multi-model ensemble results show high accuracy scores across areas similar to the best individual models in particular seasons; this is particularly evident at the shorter lead time of one month. At the longer lead time, the contributing models have similar regions of skill. Overall, the multi-models are seen to be more accurate than each of the individual models.

The contribution of P24 to the multi-model ensemble MME(b), results in higher accuracy in JJA and SON, compared to MME(a). In MAM, MME(b) is more accurate in the south east of Australia, while MME(a) is more accurate in the north-west. This accuracy in the south east comes from the strength of P24 in this region. The accuracy of MME(b) in the west in JJA is due to the high accuracy of P24 and ECMWF in this region. In SON, the strength of all of the models results in high accuracy of MME(b) at both lead times. In DJF, the two multi-models have a similar lack of accuracy, except in the west, which is a strength of the European models, as seen in MME(a). The multi-model ensemble is beaten by the UKMO individual model in terms of accuracy in DJF. However, the multi-model is more accurate than this model in the other seasons, illustrating the benefit of the more consistently accurate multi-model ensemble.

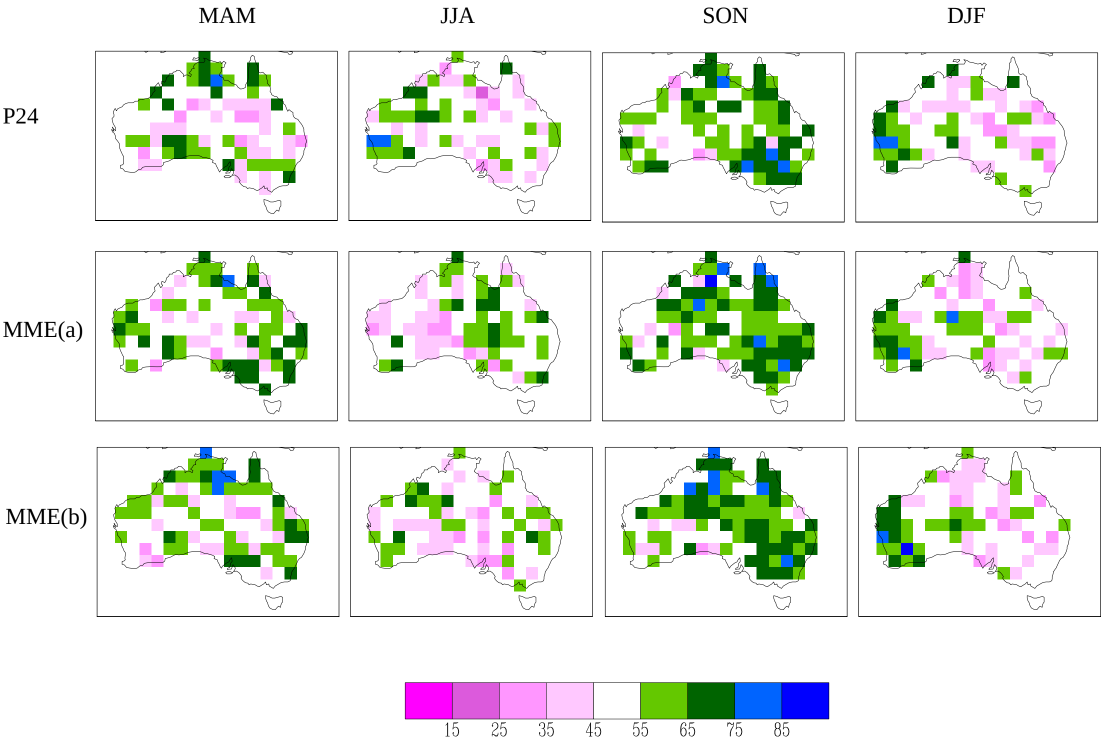

Fig.8 As for Fig. 7, for a lead time of four months.

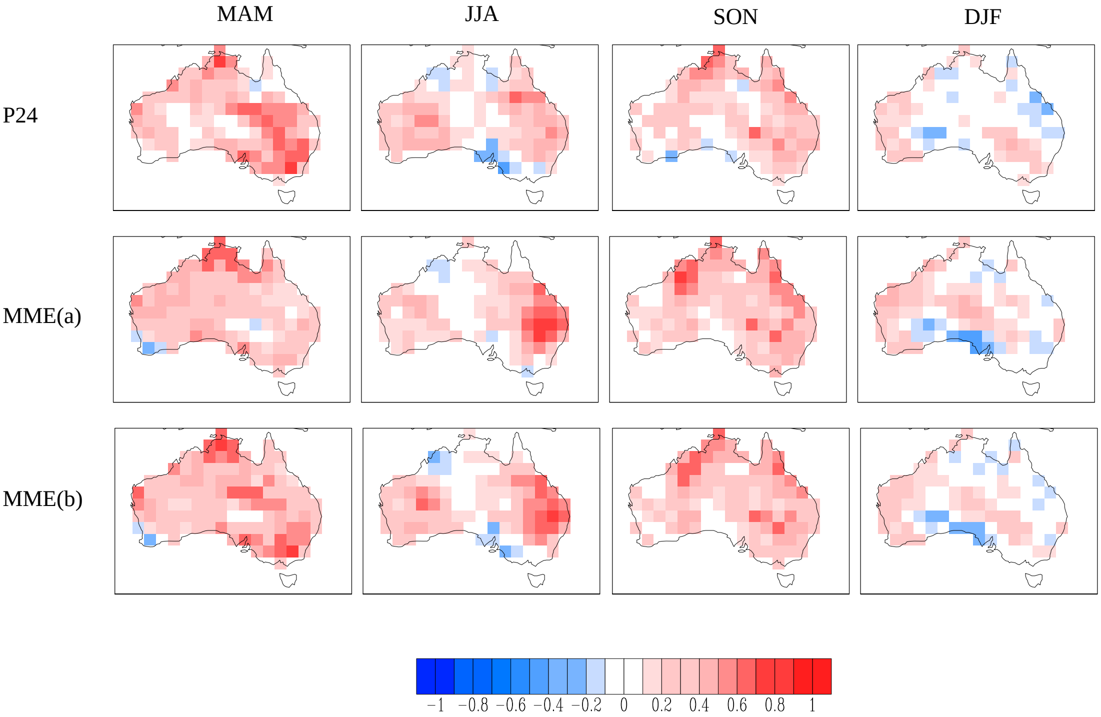

Fig.9 Correlation of ensemble mean with observations of P24 and the multi-model ensembles for a lead time of one month.

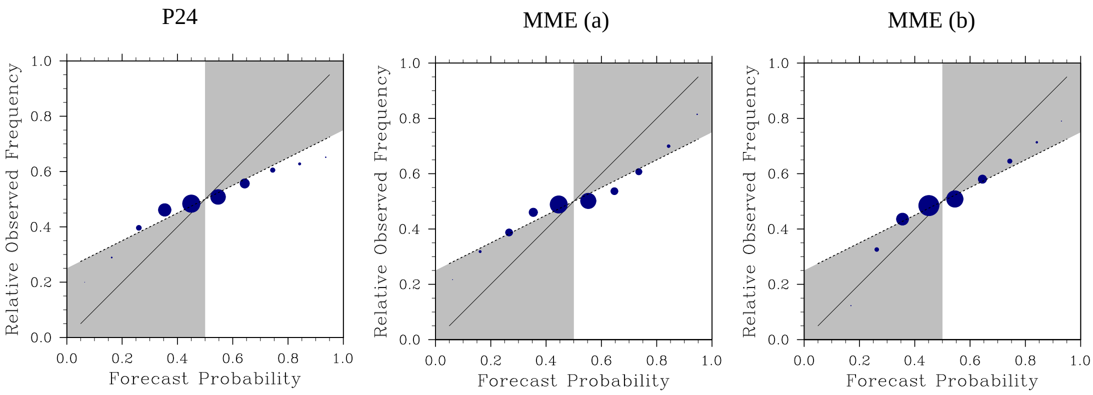

Fig.10 Attributes diagrams for P24 and the multi-model ensembles, for all four seasons and two lead times combined.

At the longer lead time of four months (Fig. 8), the multi-model ensembles are accurate in SON, particularly in the south east, and in the far west of Australia in DJF. This lack of accuracy in the other seasons is due to the lack of accuracy in the contributing models at the longer lead time. The benefit of a multi-model ensemble comes from the strengths of independent models combining together for a more consistent forecast and the cancellation of uncorrelated error. When the models have similar weaknesses, the multi-model ensemble is unable to improve the forecast accuracy. The multi-model ensemble will not be worse than the worst model at each season however, as the additional information in the other models will increase the accuracy compared to the worst model.

For the multi-model ensembles, the correlation of the ensemble mean rainfall anomaly with observations (Fig. 9) is not noticeably increased compared to the individual models (Fig. 3). The correlation is more consistently positive, with minimal regions of negative correlation, but the magnitude of the positive correlation is not increased compared to the individual models. For example, P24 has a higher correlation of predicted to observed season rainfall in the south east in MAM compared to either multi-model ensemble. This measure of deterministic accuracy does not take into account the increased spread of the model that will benefit the probabilistic forecasts by increasing the reliability.

5.2 Reliability, Resolution, Sharpness

Figure 10 shows the attributes diagrams for P24 and the two multi-model ensembles, where the forecasts from the four seasons and two lead times have been collected together, as in Fig. 4. Both multi-model ensembles are more reliable than the individual models, as seen by data pointsbeing closer to the diagonal solid line. MME(a) is similar to P24 at forecast probabilities less than 50 per cent, but more reliable at the higher forecast probabilities. MME(b) is more reliable than MME(a), but a larger proportion of the forecasts are closer to a 50 per cent probability, as was seen for P24 compared to the other individual models.

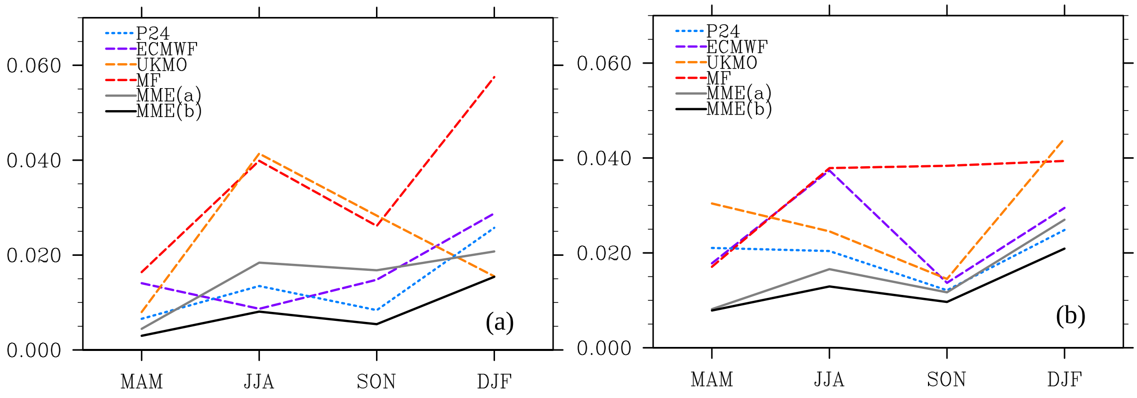

Fig.11(a) Mean reliability error as a function of season, for multi-model ensembles compared to individual models, for a lead time of one month. **(b)** For a lead time of four months.

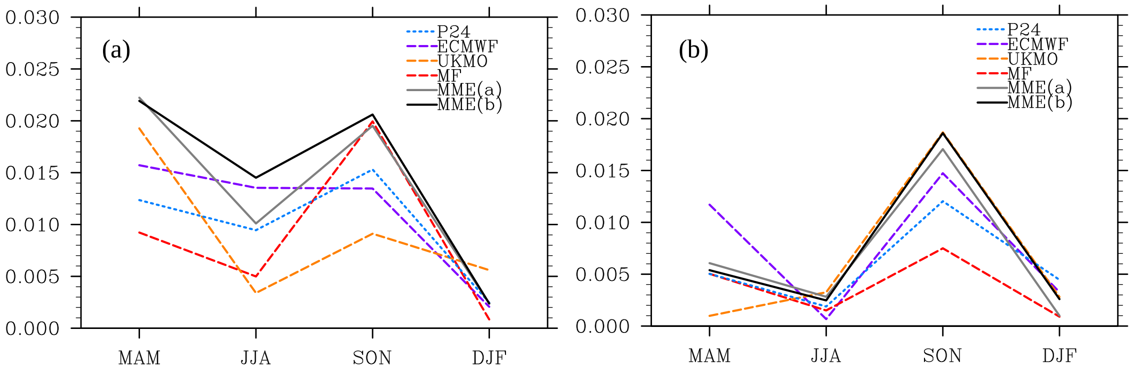

Fig.12(a) Mean resolution as a function of season, for multi-model ensembles compared to individual models, for a lead time of one month. **(b)** For a lead time of four months.

The continent averaged reliability error (Equation 5) of the multi-models compared to the individual contributing models is shown in Fig. 11 for lead times of one month and four months. The solid grey line indicates the seasonality of the reliability of MME(a). The solid black line corresponds to MME(b), which is consistently more reliable than MME(a) and all individual models for all seasons and lead times. At lead time one month, MME(a) is more reliable than the majority of the contributing individual models (dashed lines), and the individual model which is more reliable than MME(a) differs depending on the season. At the longer lead time, MME(a) is more reliable than all contributing individual models. The reliability of P24 (blue dotted line), discussed in Section 4.2, contributes to the strength of MME(b).

The average resolution over Australia of the multi-model ensembles compared to the individual models is shown in Fig. 12. The multi-model ensembles have higher resolution than most of the individual models, however they have similarly low resolution in DJF and at the longer lead time in JJA. There are a larger proportion of forecasts issued closer to 50 per cent, as seen in Fig. 10. However, the models still have high resolution due to their good reliability.

5.3 Relative Operating Characteristic

To assess whether the models are skilful and well-discriminated, the relative operating characteristic (ROC) was calculated. Discrimination depends on the distribution of forecasts given the different outcomes (Murphy and Winkler 1987). ROC curves are constructed from the hit rate plotted against the false alarm rate for an increasing probability threshold cutoff. The area under the curve is taken to be the ROC score. The hit rate is the number of occurrences of the event corresponding to forecasts up to the threshold, divided by the total number of times the event was observed. The false alarm rate is the number of non-occurrences of the event corresponding to forecasts up to the same threshold, divided by the total number of times the event did not occur.

The hit rate:

$$HR(t) = P(p \geq t \mid x = 1) \qquad (7)$$

and false alarm rate:

$$FAR(t) = P(p\geq t \mid x = 0) \qquad (8)$$

where p is the forecast probability, t is the increasing threshold and x = 1 when the event occurred and x = 0 when the event did not occur.

The ROC curve stretches from the origin (0,0) to the point (1,1). For a perfectly discriminated forecast, the area under the curve equals unity. A diagonal line and a ROC score of 0.5 indicates that hits and false-alarms occur at the same rate, which implies no skill.

An example ROC curve is shown in Fig. 13, for a single grid point over south east Australia in SON for MME(b) at a lead time of one month. The area under the curve is equal to 0.805, indicating a relatively well-discriminated forecast. Figure 14 shows the average ROC score for each model, where the area under the ROC curve has been calculated for each grid point over Australia and averaged at each season and lead time.

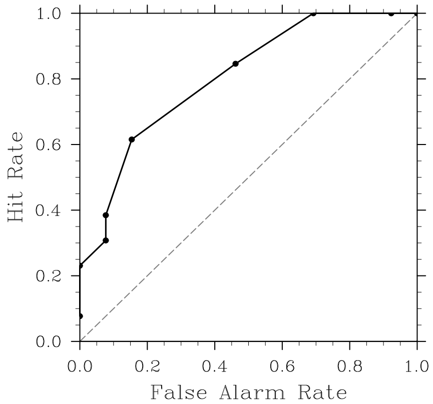

Fig.13ROC curve for a single grid point located over south east Australia at 31^{o}S 145^{o}E for MME(b) in SON at lead time one month. The grey dashed line indicates the no skill line. The area under the curve, (which is taken as the ROC score) is equal to 0.805. The solid data points indicate the hit rate and false alarm rate for the ten probability thresholds, equally spaced between 0 and 100 per cent.

The resulting seasonality of the ROC scores is quantitatively similar to that of the average resolution shown in Figs 6 and 12 (Kharin and Zwiers, 2003). These two attributes are similar, but while resolution is a measure of the distribution of the outcomes given the forecast, the

ROC score is a measure of the distribution of the forecasts given the outcome. They are therefore related measures of the quality of a forecast system. High resolution means that the outcome is well conditioned on the forecast, whereas a well-discriminated forecast will depend on the outcome (Mason and Stephenson, 2007).

The multi-model MME(b), which is more reliable than all of the contributing models, is consistently the most accurate or second most accurate model across all seasons in terms of the average ROC score. The ECMWF model in JJA and the UKMO model in DJF have the highest ROC score. MME(a) is similarly beaten by these skilful models, and by P24 in SON. The inclusion of P24 in the multi-model ensemble increases the skill in JJA and SON.

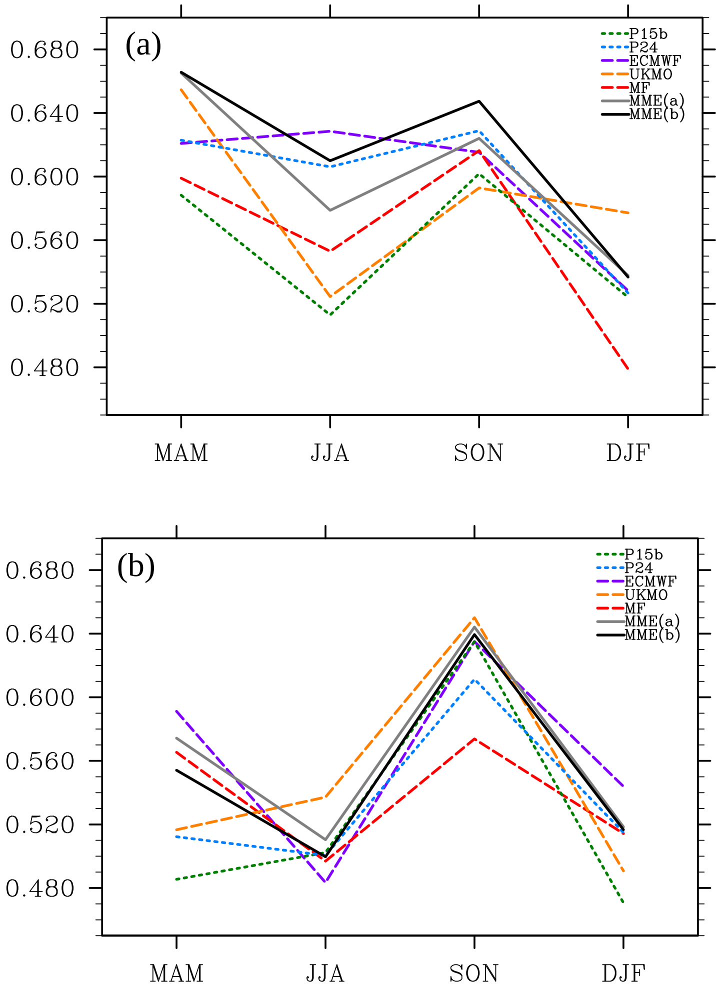

Fig.14(a) Continent averaged ROC scores as a function of season, for P15b (green dotted), P24 (blue dotted), ECMWF (purple dashed), UKMO (yellow dashed), MF (red dashed) and the multi-model ensembles (grey and black solid), for a lead time of one month. **(b)** For a lead time of four months.

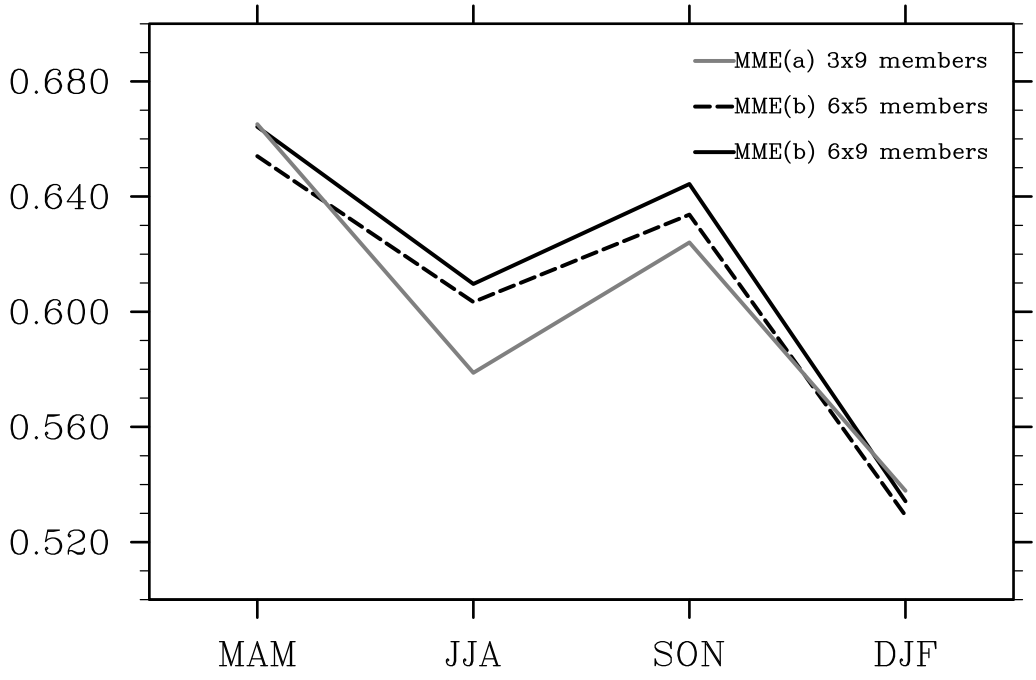

Fig.15 Continent averaged ROC scores for a lead time of one month. The grey solid line corresponds to the multi-model ensemble of ECMWF, UKMO and MF models, MME(a), where each of the three contributing models has nine members (27 total).The black solid line corresponds to the multi-model ensemble of P24, ECMWF, UKMO and MF models, MME(b), where each of the six models have nine ensemble members. The black dashed line corresponds to MME(b) calculated with only five ensemble members for each of the six models, such that the total number of ensemble members is comparable with MME(a).

At the longer lead time of four months, the skill of all models decreases, except for in SON, and the MF and ECMWF models in DJF. The individual models with the highest skill are ECMWF in MAM and DJF, and UKMO in JJA and SON, which is opposite to the shorter lead time results. The more skilful individual models at the shorter lead time show larger decreases in average ROC score at the longer lead time. MME(a) is more skilful than MME(b) at the longer lead time, due to the decrease in skill of P24 relative to the European models.

The resulting average ROC score time series is very similar to the resulting average accuracy score (not shown, area average of Figs 1, 2, 7 and 8) over the continent. Therefore we will also use the ROC score as a measure of skill in order to compare a multi-model ensemble with varying numbers of ensemble members.

Forecast skill will increase with increasing ensemble members, due to the increase in initial conditions sampled, which will decrease the uncertainty. In order to determine if the increase in skill between the two multi-model ensembles is due to the added benefit of the strengths of P24, or the increase in the number of ensemble members, we have created the equivalent of MME(b) with a smaller subset of ensemble members, comparable to MME(a). Five ensemble members are randomly selected without replacement from each of the contributing models, a multi-model probabilistic forecast is generated, and the ROC score is calculated. This is averaged over 100 Monte Carlo runs from random selections of members. This process is also repeated for nine ensemble members per model. The resulting ROC scores are shown in Fig. 15.

The solid black line in Fig. 15 shows the average ROC score for MME(b), and the dashed black line shows MME(b) with a reduced number of ensemble members. There is only a slight decrease in skill as measured by the average ROC score, despite a decrease in the total ensemble size from 54 to 30 members. The grey solid line corresponds to the MME(a) multi-model, with a total of 27 ensemble members, which is comparable to the reduced size version of MME(b) shown by the dashed black line. The increase in skill of MME(b) over MME(a) is larger than the increase seen due to the larger ensemble size. Therefore the higher skill of MME(b) is due to the reduced model error with the addition of P24 to the multi-model ensemble.

6 CONCLUSION

In this report we assessed forecasts from POAMA versions P15b and P24 in comparison to forecasts from models from ECMWF, UKMO and MF, which are archived as part of the ENSEMBLES project. The aim wasto determine how universal the lack of reliability is in dynamical forecasts of regional rainfall, in order to highlight any potential for improvement of the POAMA system. The ENSEMBLES project provides an easily accessible database of hindcast datasets for European models from ECMWF, UKMO, MF, CMCC-INGV and IFM-GEOMAR.

The systems assessed in this report show that overconfidence and lack of reliability for regional rainfall forecasts is a common problem amongst international coupled dynamical models. The individual models showed similar accuracy, but POAMA version P24 demonstrated higher reliability than the ECMWF, UKMO and MF models, as well as the older POAMA version, P15b.

Multi-model ensembles are a useful tool for utilising the strengths of a range of independent models to increase the reliability and accuracy of climate predictions. The combination of individual models into a multi-model ensemble benefits from the cancellation of uncorrelated error, increased spread and reduced model error. Two multi-model ensembles were assessed for accuracy and reliability compared to the contributing forecasts. Both multi-model ensembles included forecasts from ECMWF, UKMO and MF. The multi-model ensemble which also included P24 showed consistently higher reliability and ROC scores than the MME without P24. This increase in skill of the multi-model ensemble with addition of POAMA was shown to be due to reduced model error rather than an increase in ensemble size. These results indicate that there is benefit in adding POAMA version P24 to the available EUROSIP operational models from ECMWF, MF and UKMO, to increase the reliability and consistency of accurate regional rainfall forecasts.

Therefore, due to the clear need for improved reliability and more accurate seasonal rainfall forecasts for hydrological applications, we recommend further investigation of adopting an operational multi-model ensemble combining P24 with available European datasets.

REFERENCES

Alessandri, A., Borrelli, A., Masina, S., Cherchi, A., Gualdi, S., Navarra, A., Di Pietro, P. and Carril, A.F.(2010).The INGV-CMCC seasonal prediction system: Improved ocean initial conditions. *Mon. Wea. Rev.* 138, 2930.

Balmaseda, M.A., Vidard, A. and Anderson, D.L.T. (2008). The ECMWF ocean analysis system: ORA-S3. *Mon. Wea. Rev.* 136, 3018.

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

Collins, M., Tett, S.F.B. and Cooper, C. (2001). The internal climate variability of HadCM3, a version of the Hadley Centre coupled model without flux adjustments.*Clim. Dyn.* 17, 61.

Collins, W.J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Hinton, T., Jones, C.D., Liddicoat, S., Martin, G., O'Connor, F., Rae, J., Senior, C., Totterdell, I., Woodward, S., Reichler, T., Kim, J. and Halloran, P. (2008). Evaluation of the HadGEM2 model.*Hadley Centre Technical Note 74*, Met. Office Hadley Centre, Exeter, UK.

Daget, N., Weaver, A.T. and Balmaseda, M.A. (2009). Ensemble estimation of background-error variances in a three-dimensional variational data assimilation system for the global ocean.*Quarterly Journal of the Royal Meteorological Society* 135, 1071.

Doblas-Reyes, F.J., Weisheimer, A., Déqué, M., Keenlyside, N., McVean, M., Murphy, J.M., Rogel, P., Smith, D. and Palmer, T.N. (2009). Addressing model uncertainty in seasonal and annual dynamical ensemble forecasts. *QJR Meteorol. Soc.* 135, 1538.

Doblas-Reyes, F.J., Déqué, M. and Piedelievre, J.-P. (2000). Multi-model spread and probabilistic seasonal forecasts in PROVOST. *QJR Meteorol. Soc.* 126, 2069.

Gordon, C., Cooper, C., Senior, C.A., Banks, H., Gregory, J.M., Johns, T.C., Mitchell, J.F.B. and Wood, R.A. (2000). The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Clim. Dyn.* 16, 147.

Hagedorn, R., Doblas-Reyes, F.J. and Palmer, T.N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. *Tellus* 57A, 219.

Hewitt, C.D. and Griggs, D.J. (2004). Ensembles-based predictions of climate changes and their impacts.*Eos*, 85, 566.

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.

Johnson, C. and Swinbank, R. (2008). Medium-range multi-model ensemble combination and calibration. *Met Office Technical Report no. 517.*

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

Keenlyside, N., Latif, M., Botzet, M. and Jungclaus, J. (2005). A coupled method for initializing El Niño Southern Oscillation forecasts using sea surface temperature. Tellus 57A, 340.

Kharin, V.V. and Zwiers, F.W. (2003). On the ROC score of probability forecasts. *J. Climate*, 16, 4145.

Large, W. and Yeager, S. (2004). Diurnal to decadal global forcing for ocean and sea ice models: The data sets and climatologies.*Technical Report TN-460 + STR,* NCAR 105pp.

Lim, E.-P., Hendon, H.H., Anderson, D.L.T., Charles, A. and Alves, O. (2010). Dynamical, statistical-dynamical and multi-model ensemble forecasts of Australian spring season rainfall.*Mon. Wea. Rev.* 139, 958.

Madec, G., Delecluse, P., Imbard, M. and Levy, C. (1998). OPA8.1 Ocean General Circulation Model reference manual. *Technical note No. 11*. LODY/IPSL. Université P. and M. Curie: Paris, France.

Mason, S.J. and Stephenson, D.B. (2007). How do we know whether seasonal climate forecasts are any good? *Seasonal Climate: Forecasting and Managing Risk*, A. Troccoli et al. Eds., Springer Academic Publishers, 265.

McBride, J.L. and Nicholls, N. (1983). Seasonal Relationships between Australian rainfall and the southern oscillation. *Mon. Wea. Rev.* 11, 1998.

Muller, W.A., Appenzeller, C., Doblas-Reyes, F.J. and Linger, M.A. (2005). A Debiased Ranked Probability Skill Score to Evaluate Probabilistic Ensemble Forecasts with Small Ensemble Sizes. *J. Climate*, 18, 1513.

Murphy, A.H. and Winkler, R.L. (1987). A general framework for forecast verification. *Mon. Wea. Rev.*115, 1330.

Palmer, T.N., Alessandri, A., Anderson, U., Cantelaube, P., Davey, M., Délécluse, P., Déqué, M., Diez, E., Doblas-Reyes, F.J., Feddersen, H., Graham, R., Gualdi, S., Guérémy, J.-F., Hagedorn, R., Hoshen, M., Keenlyside, N., Latif, M., Lazar, A., Maisonnave, E., Marletto, V., Morse, A.P., Orfila, B., Rogel, P., Terres, J.-M. and Thomson, M.C. (2004). Development of a European multimodel ensemble for seasonal-to-interannual prediction (DEMETER).*Bull.Amer Meteor. Soc.*25, 853.

Pope, V.D., Gallani, M.L., Rowntree, P.R. and Stratton, R.A. (2000). The impact of new physical parameterizations in the Hadley Centre climate model: HadAM3.*Clim. Dyn.* 16, 123.

Rajeevan, M., Unnikrishnan, C.K. and Preethi, B. (2011). Evaluation of the ENSEMBLES multi-model seasonal forecasts of Indian summer monsoon variability. *Clim.Dyn.* DOI 10.1007/s00382-011-1061-x.

Reynolds, R.W., Rayner, N.A., Smith, T.M., Stokes, D.C. and Wang, W. (2002). An Improved In Situ and Satellite SST Analysis for Climate.*J. Climate* 15, 1609.

Roeckner, E., Bäuml, G., Bonaventura, L., Brokopf, R., Esch, M., Giorgetta, M., Hagemann, S., Kirckner, I., Komblueh, L., Manzini, E., Rhodin, A., Schlese, U., Schulzweida, U. and Tompkins, A. (2003). The atmospheric general circulation model ECHAM5. Part I: model description.*Report 349,* Max-Plank-Institut für Meteorologie, Hamburg, Germany.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Pan, H.-L., Behringer, D., Hou, Y.-T., Chuang, H.-Y., Iredell, M., Ek, M., Meng, J. and Yang, R. (2011). The NCEP Climate Forecast System Version 2.*To be submitted to the J. Climate*.

Salas Mélia, D. (2002). A global coupled sea ice-ocean model.*Ocean Modelling* 4, 137.

Stockdale, T.N., Anderson, D.L.T, Balmaseda, M.A., Doblas-Reyes, F., Ferranti, L., Mogensen, K., Palmer, T.N., Molteni, F. and Vitart, F. (2011). ECMWF seasonal forecast system 3 and its prediction of sea surface temperature.*Clim. Dyn.*,DOI 10.1007/s00382-010-0947-3

Uppala,S. M., KÅllberg,P.W., Simmons,A.J., Andrae, U., Da Costa Bechtold, V., Fiorino, M., Gibson, J.K., Haseler, J., Hernandez, A., Kelly,G.A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan,R.P., Andersson, E., Arpe, K., Balmaseda, M.A., Beljaars,A.C.M, Van De Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins,B.J., Isaksen, L., Janssen,P.A.E.M., Jenne, R., Mcnally, A.P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N.A., Saunders, R.W., Simon, P., Sterl, A., Trenberth, K.E., Untch, A., Vasiljevic, D., Viterbo, P. andWoollen, J. (2005). The ERA-40 re-analysis.*QJR Meteorol. Soc.*131, 2961.

Weigel, A.P., Liniger, M.A. and Appenzeller, C. (2007a). The discrete Brier and ranked probability skill scores.*Monthly Weather Review* 135, 118.

Weigel, A.P., Liniger, M.A. and Appenzeller, C. (2007b). Generalization of the discrete Brier and ranked probability skill scores for weighted multimodel ensemble forecasts. *Mon. Wea. Rev.* 135, 2278.

Weigel, A.P., Liniger, M.A. and Appenzeller, C. (2009). Seasonal ensemble forecasts: Are recalibrated single models better than multimodels?*Mon. Wea. Rev.* 137, 1460.

Weisheimer, A., Doblas-Reyes, F.J., Palmer, T.N., Alessandri, A., Arribas, A., Déqué, M., Keenlyside, N., MacVean, M., Navarra, A. and Rogel, P. (2009). ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions – Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs.*Geo. Phys. Res. Lett.* 36, L21711.

Wilks, D.S. (1995). Statistical methods in the atmospheric sciences. *Academic Press*, 467pp.

Yin, Y., Alves, O. and Oke, P.R. (2010). An ensemble ocean data assimilation system for seasonal prediction. Submitted to *Mon. Wea. Rev.*

Zhao, M. and Hendon, H.H. (2009). Representation and prediction for the Indian Ocean dipole in the POAMA seasonal forecast model. *QJR Meteorol. Soc.* 135, 337.

APPENDIX A

Table 5 The following fields are accessible from this directory by the folder labelled by their corresponding GRIB codes. (http://ensembles.ecmwf.int/download/ensembles/stream2/seasonal/atmospheric/monthly/.)

Field	Units	Code
Geopotential	m^2/s^2	129
Temperature	K	130
Zonal wind	m/s	131
Meridional wind	m/s	132
Specific humidity	kg/kg	133
Surface temperature (sea surface temperature over open waters, ice temperature over ice and temperature of the first soil layer over land)	K	139
Snow depth	m of water	141
Surface sensible heat flux	Ws/m^2	146
Surface latent heat flux	Ws/m^2	147
Mean sea level pressure	Pa	151
Total cloud cover	[0,1]	164
Zonal component of 10m wind	m/s	165
Meridional component of 10m wind	m/s	166
2m temperature (Note for UK Met Office this is the 1.5m temperature)	K	167
2m dewpoint temperature	K	168
Surface downward solar radiation	Ws/m^2	169
Surface downward longwave radiation	Ws/m^2	175
Surface net solar radiation	Ws/m^2	176
Surface net longwave radiation	Ws/m^2	177
Top net solar radiation	Ws/m^2	178
Top net longwave radiation	Ws/m^2	179
Moisture flux from the surface into the atmosphere or evaporation	m of water	182
2m Tmax (Note for UK Met Office this is the 1.5m temperature)	K	201
2m Tmin	K	202
Total precipitation	m of water	228
Vertically integrated volumetric soil water	m^3/m^3	229

The ocean reanalysis fields available by research centre at http://ensembles.ecmwf.int/download/ocean/.

The following web pages have information about the ocean reanalysis.
http://www.ecmwf.int/research/EU_projects/ENSEMBLES/data/

oras3_disclaimer.html	ECMWF
ifmkan_disclaimer.htm	IfM_GEOMAR
crfcan_disclaimer.htm	CERFACS
ingvan_disclaimer2.htm	INGV
metofficean_disclaimer.htm	Met Office

Table 6 The ocean reanalysis fields for the ENSEMBLES models

Field	Units	Code
Potential temperature	K	129
Salinity	PSU	130
Zonal velocity	m/s	131
Meridional velocity	m/s	132
Vertical velocity	m/s	133
Sea level	m	145
Mixed layer depth	m	148
20 C isotherm depth	m	163
Average potential temperature in the upper 300m.	K	164

.

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