18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Combining Verification with Probabilistic Forecasts for Decision Making

A. Charles[1], **D. Hudson**[1] **and O. Alves**[1]

[1] *Seasonal Prediction and Climate Variability Group,*
Centre for Australian Weather and Climate Research, Australian Bureau of Meteorology.
Email: a.charles@bom.gov.au

Abstract: Limits to predictability mean that in practice our projection of the future state of the atmosphere can only be probabilistic. Users of seasonal forecasts demand probabilistic forecasts for risk management and decision-making. They also require an indication of model skill which provides information about when they should use a forecast and when they should ignore it.

POAMA (Predictive Ocean-Atmosphere Model for Australia) is a coupled ocean-atmosphere model used to generate seasonal forecasts based on observed initial conditions. POAMA has been run operationally by the Bureau of Meteorology since 2002, with an initial focus on the prediction of tropical sea surface temperatures, for which the model has demonstrable skill. With recent upgrades to POAMA, there has been an increasing focus on developing regional atmospheric forecasts on seasonal time scales, organised in programs such as the South East Australian Climate Initiative (SEACI).

We demonstrate ways in which standard measures for the assessment of probabilistic forecast skill in meteorology can be incorporated into our presentation of POAMA forecasts, rather than presenting this information separately. Our current techniques for visualising the spatial and temporal distribution of skill often tend to be based on model diagnostics and are not always suitable or valuable for end users. The standard measures for the assessment of probabilistic forecast skill in meteorology such as the Brier and ROC scores are based on factorisations of the joint probability of forecasts and observations. We want to use these verification measures to improve the usefulness of probabilistic seasonal forecasts.

Chaos enforces a limit on our ability to use dynamical models to predict the future. Unavoidable imperfections in our assimilation of initial conditions, and the sensitiviy of atmospheric states to fluctuations that cannot practically be observed mean that we can only estimate the most probable state of the atmosphere. This uncertainty in initial state propagates through our model with time. Imperfections in the specification of model physics and the sensitivity of the model to errors smaller than numerical roundoff impose further limits on predictability.

Ensemble forecasting, widely used in weather and climate forecasting, attempts to quantify unpredictability by generating multiple realisations of a forecast based upon perturbed initial conditions. In essence the starting ensemble is a coarse-grained initial probability density, and the development of the ensemble provides a coarse-grained solution for the evolution of probability density in time. The ensemble spread as projected into the model future gives us a representation of many potential future states of the system of interest.

To enable forecast users to incorporate information about skill into their decision-making, we need to communicate the degree to which the verification of the dynamical forecast ought to change the forecast user's prior belief of the event's probability. In this sense the verification provides additional information about the conditional probabilities of the events we seek to forecast. The ability of dynamical models to provide skillful forecasts varies strongly by season and area. Where it is possible to resolve this variation we should provide this information. We demonstrate a novel way of presenting forecasts and verification results for seasonal rainfall forecasts for various Australian regions using the POAMA dynamical model.

A guiding assumption in our forecast product development is that forecast users are the experts in their business, and that what they need are probabilistic forecasts in which information about the level of certainty, resolution and reliability is inbuilt. Users are capable of more sophisticated decision making strategies than simply trusting or discarding a forecast. By presenting probability forecasts appropriately we can support these more advanced strategies.

Keywords: *Seasonal prediction, coupled model, probabilistic forecast*

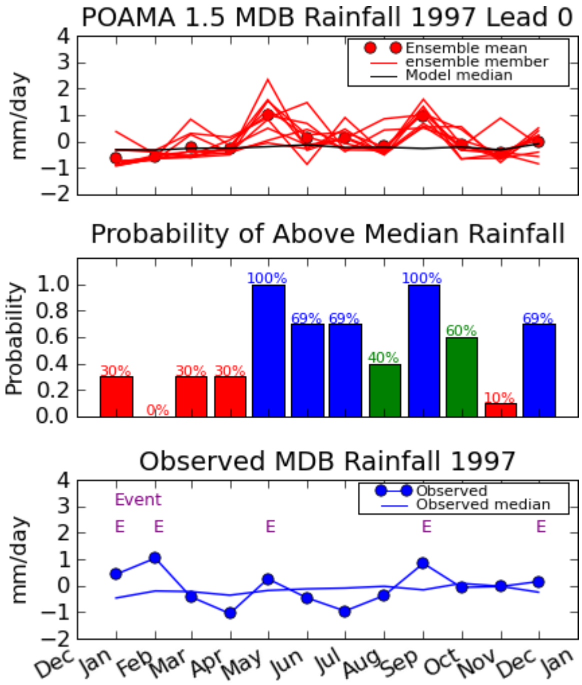

Figure 1: Conversion of an ensemble forecast into a probability forecast for above median rainfall. Top: POAMA 1.5 ensemble rainfall forecast for the year 1997. Centre: Probability of above median rainfall generated from the above ensemble. Lower: National Climate Centre Analysis for 1997. E corresponds to months when the rainfall was above median.

1. PROBABILISTIC SEASONAL FORECASTS

POAMA is a dynamical coupled ocean-atmosphere general circulation model, composed of the Bureau Atmospheric Model running at T47 spectral resolution (72 by 144 effective grid points) with 17 vertical levels and the ACOM2 ocean model. Ocean-atmosphere coupling is handled by OASIS, ocean data assimilation by an Optimum Interpolation system and land atmosphere initialisation by the ALI scheme (Alves et al., 2003); (Wang et al., 2008). The current operational version of POAMA is 1.5, for which a hindcast set exists consisting of one ten member ensemble run out to nine months from the first of each month in the period 1980 to 2006. As part of the SEACI project we provide experimental probabilistic forecasts of spatially and monthly averaged rainfall anomalies over the Murray Darling Basin (MDB) shown in figure 2.

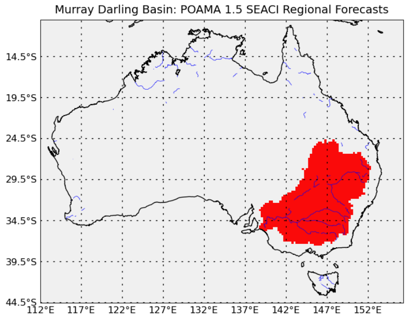

Figure 2: The Murray Darling Basin region over which monthly rainfall was averaged.

To transform POAMA's ensemble forecast into a probabilistic forecast we define one or more event thresholds, then take the fraction of ensemble members above this threshold as the probability forecast. In this study we examine the event of above median average monthly rainfall over the MDB region. Figure 1 shows the POAMA hindcast ensemble for the year 1997 and its conversion to a probabilistic forecast of above median rainfall.

Standard practice for verifying ensemble forecasts is to sort them into as many probability bins as there are ensemble members. A limitation of the POAMA hindcasts used in this study is the small size of the hindcast set available for verification. A small number of forecast verification pairs in any particular bin reduces the statistical significance of the results markedly. To mitigate this we use larger probability bins, at the expense of potential forecast resolution and sharpness (Doblas-Reyes et al., 2008). Choosing three bins for the probability of rainfall exceeding the climatological median provides an adequate sample size for most months and the probability bins translate nicely into forecasts of a low, medium and high probability of an above median rainfall event.

2. VERIFICATION SCORES

The binned forecasts were verified against Australian rainfall data from the National Climate Centre's (NCC) gridded atmospheric data set. For our three forecast bins, the contingency table summarising the forecast-verification set has the form shown in table 1, with probability bins F_1, F_2, F_3, counts of observed events o_1, o_2, o_3 and counts of non-events n_1, n_2, n_3 over each forecast. The contingency table contains all the information required to generate a standard set of verification scores (Jolliffe and Stephenson, 2003). Table 2 shows verification data for the MDB rainfall forecasts described above for all months in the hindcast period.

Table 1: Contingency table for a binary event and three forecast levels.

Forecast	Observed Events	Observed Non-events
F_1	o_1	n_1
F_2	o_2	n_2
F_3	o_3	n_3

The joint distribution of the forecasts in one bin F_i and observed events E is

$$p(F_i, Event) = \frac{o_i}{N} \tag{1}$$

where the total number of forecast-verification pairs is

$$N = o1 + o2 + o3 + n1 + n2 + n3. \tag{2}$$

The calibration-refinement factorisation of the joint distribution for a particular forecast bin

$$p(F_i, E) = p(E|F_i)p(F_i) \tag{3}$$

is composed of two factors: the true positive ratio $p(E|F_i)$ and the marginal frequency $p(F_i)$ where

$$p(E|F_i) = \frac{o_i}{(o_i + n_i)} \tag{4}$$

$$p(F_i) = \frac{(o_i + n_i)}{N}. \tag{5}$$

$p(E|F)$ is the conditional probability of the event given this particular forecast. This is the new estimate of the probability of the event of above median rainfall based on the information from the forecast and its verification

Table 2: POAMA MDB spatial average precipitation forecasts verified against NCC analysis.

Probability Bin	Events	Non-events	Total
0.00-0.32 (Low)	46	88	134
0.33-0.66 (Medium)	26	35	61
0.67-1.00 (High)	91	38	129

Table 3: Calibration function for POAMA forecasts of above median rainfall , computed using data in table 2 with 90% confidence interval.

Forecast	$p(E\|F)$	90% Confidence Interval
Low	0.34	0.26 - 0.42
Medium	0.43	0.30 - 0.54
High	0.71	0.62 - 0.78

(Murphy and Winkler, 1987). Table 3 gives the true positive ratio with 90% confidence interval for our POAMA forecasts.

As the calibration distribution in each bin is a Bernouilli distribution, confidence intervals can be generated for the forecasts, and for a set of perfect forecasts by a permutation counting method. The permutation 90% confidence interval for $p(E|F)$ is shown in table 3. We obtained similar confidence intervals using alternative methods including a bayesian technique and a bootstrap resample of the population of each bin (Brocker and Smith, 2007); (Mason, 2008).

Reliability diagrams are plots of the true positive ratio (also known as the calibration function, observed relative frequency, likelihood and hit rate) against the mean probability of the forecasts in each bin. Reliability diagrams are used to assess the degree to which the model forecast probabilities agree with the observed frequencies (Jolliffe, 2007), shown in figure 3 with the confidence intervals described above.

3. USING THE VERIFICATION FOR DECISION MAKING

A simple cost-loss model provides a framework to begin to quantify the potential value of forecasts (Jolliffe and Stephenson, 2003). In the simple binary event case, a failure to protect with cost C against an event results in a loss L. In this framework it only makes sense to take action given the probability of the event p if $p > \frac{C}{L}$. To make an optimal decision using this framework, the actual or best estimate probability of the event is needed. Given information about climatology, a model and its verification, the calibrated model probability $p(E|F)$ provides this best estimate.

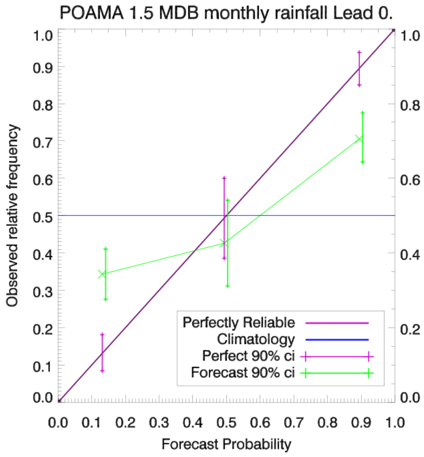

Figure 3: Reliability diagram for POAMA 1.5 Murray Darling Basin Average Monthly rainfall, for all months. The green interval is the forecast 90% confidence interval (ci), the purple interval is the 90% confidence interval for perfect forecasts with the same sample size.

In order to make rational decisions based on quantifiable costs, losses and probabilities the end user needs the calibrated forecast probabilities. The end user of the forecast knows what their costs and losses are for each contingency. If they know the calibrated forecast probabilities, with reliable confidence intervals, they are in a position to use these probabilities to determine the optimum course of action to follow for their unique cost function.

Usually skill for coupled models is presented as correlation plots, rms error plots, and sometimes Brier skill scores for probabilistic forecasts. While these scores are useful for model diagnostics, and can quantify potential forecast value, it is not obvious how users who need to make decisions based on forecasts should convert these measures into new estimates of probability.

The above model can be extended to more sophisticated decisions based on event probability thresholds, detailing different actions to be taken at different thresholds depending on the users attitude to risk. We present a hypothetical example of an agriculturalist making a decision about whether to apply additional fertilizer, at a cost, with a potential payoff depending on the probability of expected rainfall being above median (Table 4).

Table 4: Example probability thresholds for a decision about whether to apply no fertilizer, a normal amount, or a maximum amount to take advantage of expected rain.

Rainfall Event Probability	Action	Rainfall	No Rainfall
0-20%	No fertilizer	Missed profit	Minimal loss
20-70%	Normal fertilizer	Normal profit	Moderate loss
70-100%	Maximum fertilizer	Bumper crop	Greatest loss

Table 5: Application of POAMA MDB rainfall forecasts to decision table, showing the forecast category, corresponding probability interval and decision table mapping.

Forecast Probability	Action	Rainfall	No Rainfall
Low (26-42%)	Normal fertilizer	Normal profit	Moderate loss
Medium (30-54%)	Normal fertilizer	Normal profit	Moderate loss
High (62-78%)	Maximum fertilizer	Bumper crop	Greatest loss

In this example 20% rainfall probability is the threshold at which the cost of applying fertilizer is less than the expected payoff (Table 4).

The decision thresholds in table 4 provide a way of mapping from a given forecast to an action, shown in table 5. Using the true positive ratio we calculated for our sample rainfall forecasts in table 3, the decision-maker would find that the calibrated ‘low probability’ forecasts from POAMA are not sufficient to justify the ‘no fertilizer’ action, because the observed frequency of above median rainfall events is above the 20% threshold.

4. PRESENTING THE FORECAST AND VERIFICATION AS A COMBINED FORECAST

There are several ways to present information about the calibration. The actual contingency table (table 2) has the advantage of containing almost all the usable information, but the disadvantage of requiring a knowledge of verification methods to translate it into usable probabilities.

A plot of the actual ensemble of past forecasts (figure 1) allows users to eyeball the agreement and spread between forecasts and observations. However it provides no quantitative information about how much credibility to assign to the forecasts.

The reliability diagram (figure 3) provides the calibrated probability, but it is not intuitive to read. Rather than train end users to read reliability diagrams, a simple pie chart that presents the relative probabilities and level of certainty seems clearer. Figure 4 shows visually how the model forecast adjusts our estimated probabilities, and what the confidence intervals based on the size of the sample are. For each forecast category we show the prior climatological probability of the event and the updated probabilities, with 90% credible intervals for each forecast category. This plot is designed to communicate to end users how much the forecast ought to affect their estimate of the event’s probability.

Figure 5 shows the posterior probabilities for rainfall forecasts with three months lead (the fourth simulated month in the forecast). By the fourth month of the forecasts we see that the confidence intervals for the ‘high’ and ‘low’ forecasts have widened, such that the forecast of low probability is barely distinguishable from a 50/50 climatological forecast, and the 90% confidence interval for the high probability forecast almost includes the 50% climatological probability.

Coupled model skill varies strongly by month. We would like to make information about this available to users, but using the methods in this paper we do not have sufficient resolution. Table 6 shows the contingency table and true positive ratio for April forecasts. The true positive ratio suggests that the April forecasts have reasonable skill and that we ought to take the forecast of a high probability of above median rainfall as increased from a 50:50 climatological odds to 9:2 in favour of the event. Unfortunately the small sample size in each probability bin results in very large credible intervals for most months as shown in figure 6.

This inability to resolve the seasonal dependence of skill with sufficient accuracy based on this verification data is frustrating, because while we can say with confidence that useful skill exists at short lead time over

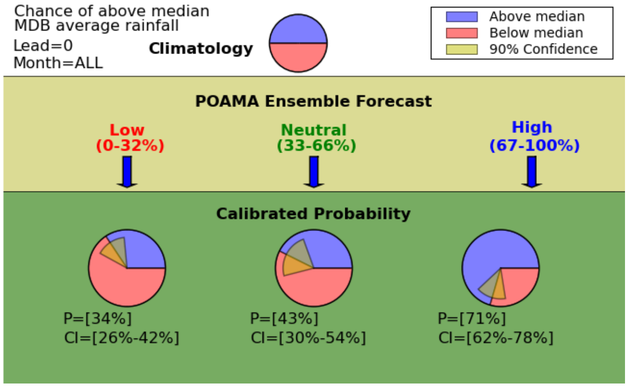

Figure 4: Posterior probability based on 3 bin POAMA forecast at 0 months lead.

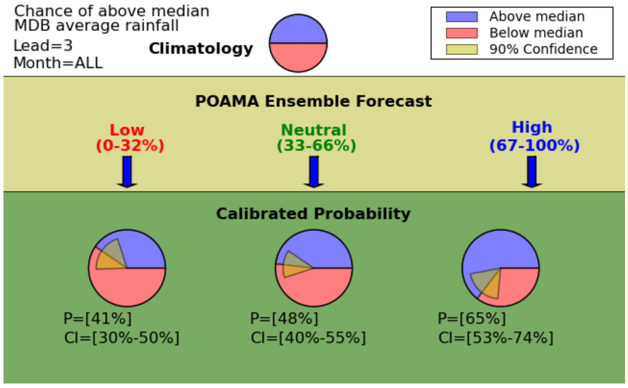

Figure 5: Posterior probability based on 3 bin POAMA forecast at 3 months lead.

Table 6: True postive ratio for POAMA MDB April rainfall

Probability Bin	$\mathbf{p(E\vert F)}$	**Events**	**Non-events**
0.00 - 0.32	0.23	03	10
0.33 - 0.66	0.33	01	02
0.67 - 1.00	0.82	09	02

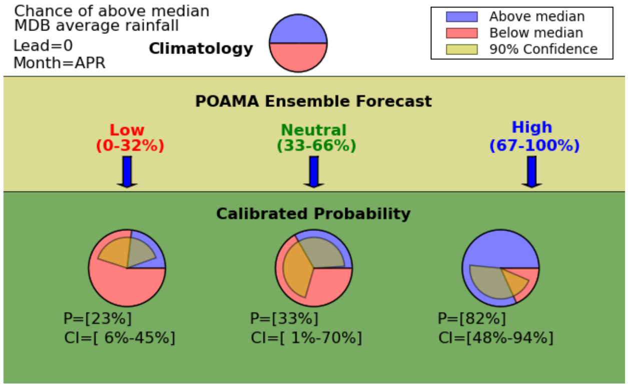

Figure 6: Posterior probability of above median rainfall for April forecasts.

all months taken together, and we know that it is likely this skill is not evenly distributed, with a handful of exceptions we cannot reliably sort the good months from the bad months. This same problem will affect the significance of attempts to calibrate forecasts for individual grid points in this manner. More sophisticated methods of calibration based on spatial patterns may mitigate this (Lim et al., 2007). Other techniques for pooling data (for example from adjacent months) may also help.

5. CONCLUSION

Simple calibration of seasonal forecasts can greatly increase the utility of seasonal forecasts for decision makers. The small size of our ensemble and hindcast period has restricted us to adding only one degree of freedom to the classic dichotomous forecast of a binary event.

More work is needed into the effect the significance level and size of the confidence interval have on the choice of optimum probability to use for decision making. Theoretical work or modelling could determine optimum forecasts for selected decision making cost functions.

The wide credible intervals around our estimate of skill by month are troubling, because we know that skill varies strongly by month but are unable to quantify this adequately for these forecasts. More research is needed into ways of pooling forecast-verification pairs in order to increase confidence. By increasing our sample size, possibly by aggregating forecasts at different locations and times we will be able to reduce the size of our confidence intervals.

A remaining application question is how the credible range should affect the decison. The wider the interval, the less evidence exists that the forecast probability corresponds to a repeatable relationship between model and reality. Decision makers may prefer to use climatological probabilities unless the evidence is above a given threshold to a given level of credibility.

The experimental forecast products we have developed will be provided for research use on the POAMA website (http://poama.bom.gov.au/).

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.

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

Doblas-Reyes, F. J., Coelho, C. A. S., and Stephenson, D. B. (2008). How much does simplification of probability forecasts reduce forecast quality? *Meteorological Applications*, 15(1):155–162.

Jolliffe, I. T. (2007). Uncertainty and inference for verification measures. *Weather and Forecasting*, 22:637.

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

Lim, E. P., Hendon, H., and Alves, O. (2007). Seasonal forecast of the tropical Indo-Pacific SST and australian rainfall SEACI tecnical report.

Mason, S. J. (2008). Understanding forecast verification statistics. *Meteorological Applications*, 15(1):31–40.

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

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