Improving Intraseasonal Prediction with a New Ensemble Generation Strategy

DEBRA HUDSON, ANDREW G. MARSHALL, YONGHONG YIN, OSCAR ALVES, AND HARRY H. HENDON
Centre for Australian Weather and Climate Research,* Melbourne, Victoria, Australia

(Manuscript received 14 February 2013, in final form 11 July 2013)

ABSTRACT

The Australian Bureau of Meteorology has recently enhanced its capability to make coupled model forecasts of intraseasonal climate variations. The Predictive Ocean Atmosphere Model for Australia (POAMA, version 2) seasonal prediction forecast system in operations prior to March 2013, designated P2-S, was not designed for intraseasonal forecasting and has deficiencies in this regard. Most notably, the forecasts were only initialized on the 1st and 15th of each month, and the growth of the ensemble spread in the first 30 days of the forecasts was too slow to be useful on intraseasonal time scales. These deficiencies have been addressed in a system upgrade by initializing more often and through enhancements to the ensemble generation. The new ensemble generation scheme is based on a coupled-breeding approach and produces an ensemble of perturbed atmosphere and ocean states for initializing the forecasts. This scheme impacts favorably on the forecast skill of Australian rainfall and temperature compared to P2-S and its predecessor (version 1.5). In POAMA-1.5 the ensemble was produced using time-lagged atmospheric initial conditions but with unperturbed ocean initial conditions. P2-S used an ensemble of perturbed ocean initial conditions but only a single atmospheric initial condition. The improvement in forecast performance using the coupled-breeding approach is primarily reflected in improved reliability in the first month of the forecasts, but there is also higher skill in predicting important drivers of intraseasonal climate variability, namely the Madden–Julian oscillation and southern annular mode. The results illustrate the importance of having an optimal ensemble generation strategy.

1. Introduction

Dynamical model prediction of intraseasonal climate is a developing field worldwide, exhibiting a growing research and operational focus (e.g., Toth et al. 2007; Gottschalck et al. 2008; Vitart et al. 2008; Brunet et al. 2010; Rashid et al. 2010; Hudson et al. 2011a,c; Marshall et al. 2011a,b,c; Vitart and Molteni 2011). Traditionally, it has been a difficult time scale for which to provide skillful forecasts, since beyond a week the system has lost much of the information from the atmospheric initial conditions (upon which weather forecasts rely), and the variability forced by slow variations of boundary conditions (e.g., tropical sea surface temperatures) usually does not appear above the noise until after 1-month lead time. However, over the past few years, with improvements to numerical prediction models, ensemble prediction techniques, and initialization, there have been good indications of skill at regional levels for forecasts in this prediction gap of beyond 1 week but shorter than a season (e.g., Vitart et al. 2008; Hudson et al. 2011a). Forecast models now also have improved representation and prediction of slower internal processes such as the tropical Madden–Julian oscillation (MJO) (e.g., Rashid et al. 2010; Marshall et al. 2011a; Vitart and Molteni 2011) that can contribute to predictability at lead times longer than those of midlatitude weather events but shorter than a season. This potential for intraseasonal prediction has recently been recognized by the World Meteorological Organization through the implementation of a new subseasonal prediction project (http://www.wmo.int/pages/prog/arep/wrwp/new/S2S_project_main_page.html). The research project aims to improve our forecast skill and our understanding of intraseasonal climate variability, as well as fostering the improvement and rollout of operational intraseasonal forecasting systems and uptake by the...
applications community. It is a joint project between the weather (the World Weather Research Program) and climate (the World Climate Research Program) research communities. This effort underscores the challenges of addressing this time scale, which is influenced by short-lived atmospheric processes as well as longer variations in, for example, the ocean, sea ice, and land.

The Australian Bureau of Meteorology (BoM) is currently engaged in developing the science and systems that could form an intraseasonal forecast service for Australia. This has been driven by strong stakeholder interest, particularly from the agricultural community (http://www.managingclimate.gov.au/publications/climag-18; http://www.grdc.com.au/Media-Centre/Ground-Cover-Supplements/Ground-Cover-Issue-87-Climate-Supplement). This intraseasonal forecasting capability is based on BoM’s Predictive Ocean Atmosphere Model for Australia (POAMA) coupled model seasonal prediction system. Hudson et al. (2011a,c), Marshall et al. (2011a,c), and Rashid et al. (2010) assessed the skill of the POAMA-1.5 system (hereafter P1.5) for making intraseasonal predictions of Australian climate and its ability to represent and predict key drivers of intraseasonal climate variability. Even though the P1.5 system was not designed for intraseasonal prediction (e.g., in real time only a single forecast was made once per day, thereby necessitating the use of a lagged ensemble with lags of N days for N members), there was promising skill for forecasts of the impending and subsequent fortnight (i.e., the mean of forecast days 1–14 and days 15–28, respectively) for rainfall and maximum temperature, particularly over eastern and southeastern Australia. Interestingly, forecast skill for fortnight 2 was found to be increased during extremes of the El Niño–Southern Oscillation (ENSO), the Indian Ocean dipole (IOD), and the southern annular mode (SAM), indicating that these slow variations of boundary forcing should be considered a source of intraseasonal climate predictability. Nonetheless, deficiencies in the P1.5 system, which was designed for seasonal climate prediction, were clearly acting to limit the capability of making intraseasonal forecasts.

In the upgrade of POAMA to version 2, the most significant changes included a more advanced ocean data assimilation scheme (Yin et al. 2011) and a multimodel formulation together with an increased ensemble size. Implicit in the new ocean data assimilation scheme is the generation of an ensemble of ocean states for initializing the forecasts. Although the POAMA-2 system uses a “burst” ensemble generation approach (i.e., all ensemble members initialized at the same time), thereby rectifying the need for a lagged ensemble to generate intraseasonal forecasts, each ensemble member uses the same atmospheric initial conditions. We will show in this paper that although the latter approach might be appropriate for seasonal prediction, it is inadequate for intraseasonal prediction, since by not perturbing the atmosphere at the initial time, there is insufficient spread in the first 30 days of the forecasts. As a result, a novel coupled ensemble initialization scheme has been developed for initializing the intraseasonal forecasts. This scheme is based on a breeding cycle performed with the coupled model and generates perturbed atmosphere and ocean states about the central control member that is provided by the ocean and atmosphere analyses. A breeding approach has been used before within a coupled ocean–atmosphere system, although the primary focus was to isolate ENSO coupled instabilities in the initial ensemble perturbations for seasonal to interannual time-scale forecasts (e.g., Toth and Kalnay 1996; Cai et al. 2003; Yang et al. 2006).

The aims of this paper are to document the intraseasonal skill of the new intraseasonal version of the POAMA-2 system, which we refer to as the multiweek version (P2-M), and to compare it to the skill of the seasonal version of POAMA-2 (P2-S) and the older P1.5 system. The focus will be primarily on forecast performance for the Australian region but impacts of the new POAMA ensemble generation strategy for prediction of the MJO and the SAM will also be documented. The characteristics (e.g., structure and growth rates) of the perturbations generated by the coupled-breedng technique will be addressed in a separate paper.

2. POAMA-2 intraseasonal forecast system
a. Model and data assimilation details

Dynamical intraseasonal–seasonal prediction at the Bureau of Meteorology is based on POAMA-2, which is a coupled ocean–atmosphere climate model and data assimilation system. Prior to March 2013, the POAMA-2 system was configured in two ways: one for multiweek lead times for intraseasonal predictions (P2-M, an experimental system) and one for seasonal predictions (P2-S, an operational system). The model components are identical and the two systems differed only in the manner of ensemble initialization and in forecast frequency and length. P2-S was run less often but forecasts were longer in order to target seasonal climate. P2-M was run more often and for shorter forecasts in order to target intraseasonal variability. System details of P2-M, P2-S, and P1.5 are summarized in Table 1. We note that in March 2013 the P2-M system replaced P2-S as the BoM operational system for intraseasonal to seasonal prediction.

The main modules of the coupled models in P2-M and P2-S are identical and similar to those of the older P1.5.
The atmospheric model component is the BoM’s Atmospheric Model (BAM version 3.0; Colman et al. 2005; Wang et al. 2005; Zhong et al. 2006), which has a T47 horizontal resolution (approximately a 250-km grid) and 17 levels in the vertical. This horizontal resolution, together with the grid configuration, means that Tasmania is not resolved as land in POAMA. The land surface component in POAMA is a simple bucket model for soil moisture (Manabe and Holloway 1975) and has three soil levels for temperature. The ocean model is the Australian Community Ocean Model version 2 (ACOM2; Schiller et al. 1997; Schiller et al. 2002; Oke et al. 2005) and is based on the Geophysical Fluid Dynamics Laboratory Modular Ocean Model (MOM version 2). The zonal resolution of the ocean grid is $2^\circ$, and the meridional spacing is 0.5° within 8° of the equator, increasing gradually to 1.5° near the poles. ACOM2 has 25 vertical levels, with 12 levels in the top 185 m and a maximum depth of 5 km. The atmosphere and ocean models are coupled using the Ocean Atmosphere Sea Ice Soil (OASIS) coupling software (Valcke et al. 2000).

In addition to configuring POAMA-2 for intra-seasonal prediction (described in section 2b), P2-M and P2-S have major upgrades over P1.5 in the following aspects:

- an advanced ocean data assimilation scheme (Yin et al. 2011),
- a multimodel ensemble (MME) approach (Wang et al. 2011), and
- a larger ensemble over a longer hindcast period.

Forecasts from all versions of POAMA are initialized from observed atmospheric and oceanic states. The atmosphere and land initial conditions are provided by the Atmosphere–Land Initialization (ALI) scheme for both P1.5 and P2 (Hudson et al. 2011b). ALI creates a set of atmosphere and land initial states by nudging zonal and meridional winds, temperatures, and humidity from the atmosphere model of POAMA (run prior to hindcasts or forecasts being made and forced with observed SST) toward an observationally based analysis. ALI nudges to the analyses from the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40; Uppala et al. 2005) for the period 1980–August 2002, and to the BoM’s operational global numerical weather prediction (NWP) analysis thereafter. The scheme also generates land surface initial conditions that are in balance with the atmospheric condition.

The ocean analysis system for P1.5 was based on a univariate optimum interpolation system that assimilates only in situ temperature observations (Smith et al. 1991). Due to the lack of appropriate multivariate formulations, this approach has a detrimental effect on the salinity and velocity fields of the model. To rectify these limitations, a new ocean analysis system, the POAMA Ensemble Ocean Data Assimilation System (PEODAS; Yin et al. 2011), has been developed. PEODAS is an approximate form of the ensemble Kalman filter and generates an ensemble of ocean states each day, including a central unperturbed ocean analysis. PEODAS is based on the multivariate ensemble optimum interpolation system of Oke et al. (2005), but uses covariances derived from the time-evolving model ensemble. In situ temperature and salinity observations are assimilated, and corrections to currents are generated based on the ensemble cross covariances with temperature and salinity. PEODAS has demonstrated a significant quantitative improvement in the depiction of the initial state of the upper ocean, especially in relation to salinity (Yin et al. 2011; Zhao et al. 2013b).

To address model uncertainty, POAMA-2 (both P2-S and P2-M) has adopted a pseudo-MME strategy using three different model configurations:

- P2a—modified atmospheric physics configured to use an alternative shallow convection physical parameterization
scheme that results in less climate drift compared to P1.5.
- P2b—bias-corrected version of P2a (i.e., same model version as P2a but with ocean–atmosphere fluxes corrected to reduce climatological biases), and
- P2c—standard physics options that are identical to those in P1.5.

b. Ensemble generation

The 10-member ensemble of forecasts for P1.5 was generated using lagged atmospheric initial conditions from ALI by successively initializing each member with the atmospheric analysis from 6 h earlier (i.e., the 10th member was initialized 2.25 days earlier than the first member). Each ensemble member used the identical ocean initial condition.

In contrast, POAMA-2 uses a burst ensemble (i.e., all initial conditions are valid for the same date and time; there are no lagged initial conditions). The 30-member P2-S ensemble comprises a 10-member ensemble for each model version (P2a, P2b, P2c) that is generated using 10 ocean initial conditions provided directly by PEODAS (each model version uses the same set of 10 ocean initial conditions). However, in contrast to P1.5, all ensemble members use the identical atmospheric initial condition. Not perturbing the atmosphere is acceptable for seasonal predictions, but for intraseasonal prediction there is not enough spread generated by the model in the first month of the forecast. This can be seen in Fig. 1, which shows that the ensemble spread of 500-hPa heights from P2-S at day 10 of the forecast is an order of magnitude smaller than that obtained from P1.5 (Figs. 1a and 1b). However, the spread at day 10 from P2-M (Fig. 1c) is comparable (if not slightly larger) to that obtained from P1.5. Details of the ensemble generation strategy for P2-M are described below.

The P2-M forecasts are initialized using a coupled-breeding scheme, which we refer to as the Coupled Ensemble Initialization (CEI) scheme. A schematic of the method is shown in Fig. 2. The breeding scheme produces 10 perturbed states about the central analyses provided by the PEODAS central analysis for the ocean and the ALI analysis for the atmosphere–land. During each breeding cycle the following steps are taken:

1) The 11-member ensemble (using the P2a coupled model), including one central unperturbed control and 10 perturbed members, is integrated forward 24 h from an initial time \( t \). The purpose of the central coupled model run is to generate the 24-h forecast for calculation of the perturbations and this control is initialized at time \( t \) using the central analysis (i.e., PEODAS for the ocean and ALI for the atmosphere and land).
2) At the end of the 24-h integration \( (t + 24) \), the bred perturbations (atmospheric \( u, v \), \( T, q \) and oceanic \( T, S, u, v \)) are determined as the difference between each perturbed ensemble forecast and the central control forecast.

![Fig. 1. Ensemble spread of 500-hPa heights (m) at day 10 of forecasts initialized on 1 Jul 1980–2006 for (a) P1.5, (b) P2c-S, and (c) P2c-M. All three systems use the same model (see text for details) and results are displayed for 10-member ensembles. Spread greater than 10 m is depicted by solid contours with an interval of 10 m. Spread less than 10 m is depicted by dashed contours with an interval of 2 m. The latter was done to emphasize that all spread in (b) is less than 10 m.](image-url)
3) The forecast perturbations of each ensemble member are then scaled with respect to the estimated analysis error (discussed below) and centered by subtracting the mean of the rescaled perturbations from all ensemble members. The expectation is that this limited number of perturbations will have an initial spread distribution that is similar to the analysis error variance.

4) The ensemble members are then updated by adding the perturbations to the central analysis (PEODAS for the ocean and ALI for the atmosphere–land) at time $t + 24$. These 10 coupled updated states, together with the central analysis, are used for initializing the hindcasts and forecasts.

5) Repeat steps 1–4 above.

All ensembles are cycled every 24 h for the entire 1980–2010 period. Prior to this period, there is an initial 2-week spinup, started from ocean perturbations from PEODAS. In this spinup there is no rescaling, so as to allow fast weather instabilities to saturate. CEI uses a one-sided breeding method (i.e., rather than positive–negative paired or two sided), also referred to as subtract-mean breeding (Toth and Kalnay 1997; Wang et al. 2004). Wei et al. (2008) found that experiments using one-sided breeding have performed significantly better than the paired perturbations with performance almost as high as that found with the ensemble transform technique. The better performance with one-sided breeding is related to the ensemble centering strategy. By simply removing the mean from all perturbations, the independence of these perturbations is preserved, whereas the paired centering scheme reduces the effective number of degrees of freedom of the ensemble subspace by half and may result in worse probabilistic scores. The one-sided breeding method in CEI is similar to the centering strategy used in the PEODAS ensemble setup (Yin et al. 2011).

Considering the different characteristics of the time scales of variability in the ocean and atmosphere and their strong spatial dependence, different rescaling norms that are functions of space are used for each component. Zonally averaged root-mean-square difference (RMSD) of 10-m zonal wind ($u_{10}$; only a function of latitude) and three-dimensional (3D) RMSD of ocean temperatures have been selected as the atmospheric and oceanic rescaling norms, respectively. The rescaling factor for each ensemble member is defined as the ratio of the norm calculated from the estimated analysis error variance and the norm of the forecast perturbations at each grid point. If the variance of the ensemble perturbations about the central member does not exceed the estimated analysis error variance, then no rescaling for that day is applied. The analysis error of $u_{10}$ is estimated by the RMSD of $u_{10}$ between ALI and ERA-40 (Uppala et al. 2005). The 3D analysis error of oceanic temperature is estimated by the ensemble spread from PEODAS.

The CEI scheme is closely aligned with the long-term goal of fully coupled data assimilation with POAMA. The perturbations are generated using the coupled ocean–atmosphere model and therefore the resulting atmosphere and ocean initial perturbations are consistent. These perturbations also display significant state dependence for both atmosphere and ocean components.

The 33-member P2-M ensemble comprises an 11-member ensemble for each model version (P2a, P2b, P2c). Each model version uses the same set of 11 initial conditions (i.e., the 10 perturbed and 1 central analysis provided directly by the CEI scheme).

c. Hindcast and real-time configuration

Hindcasts from P2-M were generated 3 times per month (on the 1st, 11th, and 21st) for the period 1980–2010 and the real-time system is initialized every Thursday at 0000 UTC. P2-S hindcasts were made just once per month (on the 1st) and the real-time seasonal forecasts were made on the 1st and 15th of each month. Although the P2-S system can be used for intraseasonal prediction, the new P2-M system is more ideally suited because it is initialized once per week (or once every 10 days for the hindcasts) and the atmospheric initial conditions are perturbed. To demonstrate the improved capability of the P2-M system, this paper compares intraseasonal hindcasts from P2-M to intraseasonal hindcasts from P2-S and P1.5 over the common period 1980–2006, for hindcasts initialized on the first of the month.

d. Verification methodology

Forecast skill is assessed using anomalies from the hindcast climatology. These anomalies are created by producing a lead-time-dependent ensemble mean

![Fig. 2. The CEI system.](image-url)
climatology from the hindcasts. This climatology is a function of both start date and lead time, and thus a first-order linear correction for model mean bias is made (e.g., Stockdale 1997).

The focus of this paper is on the skill of the forecast for the first and second fortnight (average of days 1–14 and 15–28, respectively) for precipitation and temperature over Australia. The BoM National Climate Centre gridded daily analyses (averaged into fortnights) of precipitation and maximum (Tmax) and minimum (Tmin) temperature are used for the verification. These gridded analyses are produced from quality-controlled station data by the application of a three-pass Barnes successive-correction analysis (Mills et al. 1997). They are available on a 0.25° grid and are averaged onto the POAMA T47 atmospheric grid (roughly 250 km). Observed anomalies are created relative to the observed climatology (using the 1980–2006 period) for direct comparison with the POAMA anomalies. The forecast quality for 500-hPa geopotential height anomalies across the extratropical Southern Hemisphere are also assessed, which provides some insight into the capability of predicting the large-scale synoptic variability of the midlatitude circulation. The National Centers for Environmental Prediction–Department of Energy (NCEP–DOE) Atmospheric Model Intercomparison Project (AMIP-II) Reanalysis (R-2) (Kanamitsu et al. 2002) of daily 500-hPa heights are used for verification.

Verification of the ensemble mean is shown in the form of correlation, which measures the linear correspondence between the ensemble mean forecast and observed anomalies, and the root-mean-square error (RMSE) of the anomalies, which provides information on errors in forecast amplitude. For probabilistic verification, focus is placed on the forecast probability of exceeding tercile thresholds. Calculation of the terciles from both the model and observations are subject to leave-one-out cross validation. The verification metrics for the probabilistic forecasts are based on contingency tables of the number of observed occurrences and non-occurrences of the event (e.g., temperature in the upper tercile) falling into predefined forecast probability bins. In this study, five probability bins are defined: 0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, and 0.8–1.0. For calculation of the tercile thresholds, anomaly data from all of the ensemble members are used. Relative operating characteristic (ROC) scores and curves, attributes diagrams, and Brier score metrics are used (e.g., Mason and Graham 1999; Mason and Graham 2002; Toth et al. 2003; Wilks 2006; Mason and Stephenson 2008). The relative quality of the Brier score is measured with the Brier skill score (BSS), which is defined as the improvement of the probabilistic forecast relative to a reference forecast, usually climatology. Positive (negative) BSS values indicate forecasts that are better (worse) than climatology. In this study we use a formulation of the BSS referred to as the debiased BSS (Müller et al. 2005; Weigel et al. 2007; Mason and Stephenson 2008). Further details of these verification methods are provided in Hudson et al. (2011a).

3. Hindcast verification

This section compares the complete ensemble of P2-M (33-member ensemble) with P2-S (30-member ensemble) and P1.5 (10-member ensemble) in order to highlight the overall benefit of the new P2-M system.

a. Geopotential height over the Southern Hemisphere

Characteristics of ensemble spread (standard deviation of the ensemble members about the ensemble mean) and ensemble mean RMSE with lead time for daily 500-hPa geopotential height anomalies over the Southern Hemisphere extratropics (20°–60°S) for the three systems are shown in Figs. 3a–c. The scores are all normalized by the observed standard deviation. For a climatological forecast, the normalized RMSE (NRMSE) would be one, and thus forecasts are typically deemed skillful for NRMSE < 1 (i.e., forecasts are skillful when the error is smaller than that from forecasting climatology). P2-M (Fig. 3c) skillfully predicts daily geopotential height anomalies out to about 12 days, compared to 8 days for P1.5 (Fig. 3a) and about 6 days for P2-S (Fig. 3b). P2-S (Fig. 3b) starts off with a similar magnitude NRMSE to P2-M, but the error in P2-S grows rapidly and instead of asymptotically approaching a climatological forecast (NRMSE = 1), it overshoots this threshold. This is due to the lack of ensemble spread in P2-S (Fig. 3b) such that the ensemble mean still has significant amplitude (less canceling out than if there was more spread). P2-M also has a smaller error than P1.5 in the first 20 days of the forecast and a larger ensemble spread. The ensemble spread in P1.5 grows more slowly compared to P2-M, indicating that lagged atmospheric initial conditions are not an optimal way to generate forecast spread. In a perfect ensemble system, over a large sample of forecasts, the ensemble spread will be equal to the RMSE of the ensemble mean for all lead times (Palmer et al. 2006; Weisheimer et al. 2011). The P2-M system has the best relationship between error and spread among the three systems. The other two systems are clearly underdispersive (the ensemble spread is smaller than the error).

Figure 3d shows the average spatial (pattern) correlation coefficient as a function of lead time for 500-hPa height anomalies over the Southern Hemisphere. There
is not much difference between the three systems over the first 5 days of the forecast, but as the lead time increases, P2-M emerges as having a slightly higher correlation so that around lead times of 10 days, P2-M is about 2 days better than P2-S and 1 day better than P1.5 (although the differences between these correlations are not statistically significant at a 10% level, calculated using Fisher’s Z transformation and \( n = 81 \)).

\( b. \) Precipitation and temperature forecasts over Australia

ROC curves and attribute diagrams for forecasts of precipitation across Australia in the upper tercile for the first and second fortnights of the forecast for all three forecast systems are shown in Fig. 4. These curves and diagrams are based on contingency tables using all land grid boxes over Australia for all forecast start months. ROC curves provide information on forecast resolution by measuring the ability of a forecast to discriminate between the occurrence and nonoccurrence of an event, in this case, rainfall falling in the upper tercile. The no-skill line on a ROC curve is the diagonal, where hit rates equal false alarm rates. A forecast system with positive skill has a curve that lies above the diagonal and bends toward the top-left corner (0.0, 1.0), such that hit rates exceed false alarm rates. The ROC curves exhibit a decline in skill from the first to the second fortnight in all three systems, although skill still prevails in the second
fortnight (Figs. 4a and 4b). The ROC curve for P2-M is better than the other two systems in the first fortnight, but there is not much difference between the three systems in the second fortnight.

The main difference between the forecast systems is evident in the attribute diagrams. An attribute diagram shows the conditional relative frequency of occurrence of an event (observed relative frequency) as a function of the forecast probability, thereby providing information on forecast reliability. The diagonal line on the diagram indicates perfect reliability and points falling in the shaded region are forecasts that contribute positively to forecast skill (as defined by the Brier skill score, thus incorporating a combined effect of reliability and resolution) (Wilks 2006). Clearly, P2-M has greatly increased reliability compared to P2-S and P1.5 (Figs. 4c and 4d). The distinction in reliability between the systems is less dramatic in the second fortnight, but in contrast to the other two systems, the forecasts from P2-M still contribute positively to the overall skill (as indicated by being in the shaded area of the diagram; Fig. 4d). All the forecast systems are overconfident (points lie below
the diagonal for high probabilities and above the diagonal for low probabilities; Figs. 4c and 4d), but this is particularly apparent for P2-S and P1.5. Overconfidence is a common situation for seasonal forecasting (Mason and Stephenson 2008). It is evident from Fig. 4 that the older P1.5 system is less overconfident and more reliable than the newer P2-S system, particularly during fortnight 1. This is essentially due to the fact that P1.5 perturbs the atmosphere initial conditions, whereas P2-S does not. The consequence of the latter is a lack of ensemble spread in P2-S (compared to P1.5) during the first 3 weeks of the forecast such that the spread does not encompass the error; the model is underdispersive (e.g., see Fig. 3). This small spread also has the effect of dramatically reducing the effective ensemble size and a reduced ensemble size is known to have a negative impact on forecast reliability (Doblas-Reyes et al. 2008).

Deviations from perfect reliability in the diagram can be due to sampling limitations rather than necessarily true deviations from reliability (e.g., Toth et al. 2003). As such, a reliability diagram is accompanied by a histogram that indicates the sample size in each probability bin. The histogram can be used to indicate the confidence associated with the result for each probability bin, as well as the sharpness of the forecast system. If all the forecasts fell in the model climatology probability bin, then the system would have no sharpness (sharpness is the tendency to forecast extreme values). The histograms accompanying the reliability diagrams for the forecast in the second fortnight (Fig. 4d) indicate that for P2-M, the forecast probabilities peak in frequency at the climatological probability. This is in contrast to the results from the first fortnight (Fig. 4c), where the frequency of forecasts peaks in the lowest probability bins. This indicates that as the forecast lead time progresses, the forecasts become less sharp (i.e., there is a tendency toward forecasts of climatology). Note that owing to the lack of ensemble spread in P2-S, forecast probabilities are still sharp during the second fortnight (Fig. 4d). This is problematic, since the associated forecasts are not reliable.

Spatial and seasonal variations in forecast skill are shown in Fig. 5 in the form of ROC areas for predicting rainfall in the upper tercile. A ROC area (area under the ROC curve) below 0.5 implies a skill level lower than climatology [statistically significant ROC areas are shaded in Fig. 5, determined using the Mann–Whitney U statistic; Mason and Graham (2002); Wilks (2006)]. ROC areas in the first fortnight are high across most of the country (not shown), but there is a pronounced degradation in skill in the second fortnight. Most of the skill in the first fortnight comes from the first week of the forecast (not

![Fig. 5. ROC area of the probability that precipitation averaged over the second fortnight (days 15–28) is in the upper tercile for each of the four seasons for (top) P1.5, (middle) P2-S, and (bottom) P2-M. ROC areas significant at the 5% significance level are shaded (Mann–Whitney U test) and the contour interval is 0.05.](image-url)
shown). Even in the first fortnight when all three systems have their highest skill, P2-M is a clear improvement over P1.5 and especially over P2-S (not shown). In the second fortnight (Fig. 5), P2-M is again generally more skillful than P1.5 and P2-S, although the areas of skill now contract and are generally highest in the east in the austral winter and spring. This is where the impacts of ENSO and the IOD [i.e., SST anomaly differences between the western and eastern tropical Indian Ocean; Saji et al. (1999)] on rainfall are strongest and express themselves even as early as the second fortnight of the forecast (Hudson et al. 2011a).

As for rainfall, ROC curves, attribute curves, and spatial ROC area plots for Tmax (above the upper tercile) and Tmin (below the lower tercile) are shown in Figs. 6–9. Forecasts of Tmax in the first fortnight from P2-S and P1.5 are significantly less reliable and contribute less to the overall skill than P2-M (for the most part the P2-S and P1.5 points fall outside the skillful shaded area; Fig. 6c). In the second fortnight, Tmax forecasts from P2-S and P1.5 are unreliable, but forecasts from P2-M remain somewhat reliable (Fig. 6d). In particular, the reliability curve for P2-S in the second fortnight represents a system that has very little resolution: the conditional outcome is virtually the same regardless of the forecasted probability; that is, the forecasts are unable to resolve times when upper tercile Tmax is more or less likely than the overall climatological probability (Fig. 6d). The same can be said of all the systems for Tmin in the second fortnight; they are
all unreliable, exhibit very little resolution, and do not contribute positively to the overall skill (Fig. 8d). In contrast, for the first fortnight the forecasts from P2-M do contribute positively to overall skill and are more reliable than P2-S and P1.5 (Fig. 8c). ROC areas for Tmax and Tmin exhibit similar dropoffs in skill from the first to second fortnights, as did rainfall. However, Tmax forecasts are generally more skillful (higher ROC scores) than those for rainfall, especially for the first fortnight (Figs. 4 and 6). As for rainfall, P2-M is again generally more skillful than P1.5 and P2-S for Tmax over Australia (Figs. 6 and 7) and in the second fortnight this skill is highest over eastern Australia during spring (September–November, SON; see Fig. 7). For forecasts of the second fortnight, not only is Tmin the least reliable of the three variables (Fig. 8d) but it also tends to have the lowest ROC areas (Fig. 9). The reasons why the model has less skill for Tmin compared to Tmax are unclear. As suggested in Hudson et al. (2011a), it may be related to model error, but it is possible that it is a reflection of the reduced predictability for Tmin. In an examination of the relationship between ENSO and Australian land surface temperature, Jones and Trewin (2000) found that statistically significant correlations between the Southern Oscillation index (SOI) and seasonal mean Tmin were less widespread than those for Tmax. Similarly, Hendon et al. (2007) found that the relationship of the SAM with daily Tmin over Australia was weaker than the relationship with daily Tmax.

Interestingly, the greater reliability in the P2-M system compared to P2-S and P1.5 is still evident on seasonal time scales. Figure 10 shows the attributes diagrams from each system for rainfall, Tmax, and Tmin over Australia for the first season (zero lead) of the forecast. P2-M has increased reliability, resolution, and skill compared to P1.5 and P2-S for rainfall and Tmax, such that the curve clearly lies within the skillful region for P2-M. For Tmin the forecasts from all three systems are unreliable, show little resolution, and do not contribute positively to skill. Beyond the first season, the differences in reliability between P2-M and P2-S are small (if any), suggesting that the improved reliability and skill in rainfall and Tmax are coming primarily from the first month of the forecast. Also of interest, is that in contrast to intra-seasonal time scales, beyond the zero-lead first season, P2-S becomes much more reliable for Tmax (not shown) and rainfall (Wang et al. 2011) than is shown in Figs. 4, 6, and 10, and is noticeably more reliable than P1.5 (in contrast to Figs. 4, 6, and 10). The greater reliability of both P2-S and P2-M at seasonal time scales compared to P1.5 is due more to the use of a multimodel combination rather than a larger ensemble size (Wang et al. 2011).

![Fig. 7. As in Fig. 5, but for maximum temperature above the upper tercile.](image-url)
c. The MJO and SAM

We also demonstrate a positive impact of the new ensemble generation strategy for the prediction of some of the key components of the large-scale atmospheric circulation that support intraseasonal predictability. Previous work has investigated the ability of P1.5 to predict the MJO (Rashid et al. 2010; Marshall et al. 2011a) and the SAM (Marshall et al. 2011c), both important drivers of intraseasonal climate variability over Australia (Hendon et al. 2007; Risbey et al. 2009; Wheeler et al. 2009). Marshall et al. (2011b) provided an update to this work by comparing the skill in predicting these indices by P2-M with that from P1.5. Here, we add to this work by including P2-S in the comparison in order to investigate the impact of the different ensemble generation strategies on the prediction of the indices [note: the hindcast years used in Marshall et al. (2011b) differ from the present study, the former using 1989–2006].

The MJO is a prominent mode of tropical intraseasonal variability consisting of large-scale coupled patterns in atmospheric circulation and deep convection that propagate eastward over the equatorial Indian and western Pacific Oceans. The state of the MJO is described using the bivariate Real-time Multivariate MJO (RMM) index (Wheeler and Hendon 2004). The RMM index is calculated from a combined empirical orthogonal function (EOF) analysis of 15°S–15°N-averaged outgoing longwave radiation (OLR) and 850- and 200-hPa zonal wind anomalies. The predicted RMM index for POAMA
is computed by projecting each ensemble member’s predicted equatorially averaged OLR and zonal wind anomalies onto the observed EOF pair [calculated from the NCEP–National Center for Atmospheric Research (NCAR) reanalysis; Kalnay et al. (1996)]. The forecasts are then verified using a bivariate correlation and bivariate RMSE. For more details of this verification technique, see Rashid et al. (2010). Figure 11 shows the forecast skill for the bivariate RMSE and correlation using the ensemble mean. P2-M has improved forecasts of the MJO compared to P1.5 and P2-S. P2-S has the largest error (Fig. 11a) and lowest correlation (Fig. 11b) of the three systems over the 35 days shown. In terms of the bivariate RMSE, forecast skill improves by about 1 week lead time for P2-M compared to P1.5 for lead times beyond 2 weeks (e.g., the error at 15 days in P1.5 is equivalent to the error at 21 days in P2-M; Fig. 11a). The bivariate RMSE of P2-M reaches the RMSE obtained for a climatological forecast after 33 days (a climatological forecast of the RMM index has an RMSE equal of the MJO compared to P1.5 and P2-S. P2-S has the largest error (Fig. 11a) and lowest correlation (Fig. 11b) of the three systems over the 35 days shown. In terms of the bivariate RMSE, forecast skill improves by about 1 week lead time for P2-M compared to P1.5 for lead times beyond 2 weeks (e.g., the error at 15 days in P1.5 is equivalent to the error at 21 days in P2-M; Fig. 11a). The bivariate RMSE of P2-M reaches the RMSE obtained for a climatological forecast after 33 days (a climatological forecast of the RMM index has an RMSE equal of the MJO compared to P1.5 and P2-S. P2-S has the largest error (Fig. 11a) and lowest correlation (Fig. 11b) of the three systems over the 35 days shown. In terms of the bivariate RMSE, forecast skill improves by about 1 week lead time for P2-M compared to P1.5 for lead times beyond 2 weeks (e.g., the error at 15 days in P1.5 is equivalent to the error at 21 days in P2-M; Fig. 11a). The bivariate RMSE of P2-M reaches the RMSE obtained for a climatological forecast after 33 days (a climatological forecast of the RMM index has an RMSE equal of the MJO compared to P1.5 and P2-S. P2-S has the largest error (Fig. 11a) and lowest correlation (Fig. 11b) of the three systems over the 35 days shown. In terms of the bivariate RMSE, forecast skill improves by about 1 week lead time for P2-M compared to P1.5 for lead times beyond 2 weeks (e.g., the error at 15 days in P1.5 is equivalent to the error at 21 days in P2-M; Fig. 11a). The bivariate RMSE of P2-M reaches the RMSE obtained for a climatological forecast after 33 days (a climatological forecast of the RMM index has an RMSE equal of the MJO compared to P1.5 and P2-S. P2-S has the largest error (Fig. 11a) and lowest correlation (Fig. 11b) of the three systems over the 35 days shown. In terms of the bivariate RMSE, forecast skill improves by about 1 week lead time for P2-M compared to P1.5 for lead times beyond 2 weeks (e.g., the error at 15 days in P1.5 is equivalent to the error at 21 days in P2-M; Fig. 11a). The bivariate RMSE of P2-M reaches the RMSE obtained for a climatological forecast after 33 days (a climatological forecast of the RMM index has an RMSE equal
Figure 11a also shows that the bivariate ensemble spread is always smaller than the error in all three systems, indicative of underdispersed systems. However, P2-M is a marked improvement over the other two systems, suggesting that the new ensemble generation technique had a positive impact on increasing the ensemble spread, but not at the expense of increasing the RMSE. As for the 500-hPa heights, the RMSE of P2-S overshoots the expected climatological threshold due to a lack of ensemble spread (Fig. 11a).

If we consider a correlation of 0.5 as the threshold for defining skillful forecasts of the MJO (e.g., Kang and Kim 2010; Rashid et al. 2010; Arribas et al. 2011; Vitart and Molteni 2011), then the correlation of the bivariate RMM index drops to 0.5 at days 17, 20, and 22.5 in P2-S, P1.5, and P2-M, respectively (Fig. 11b). This result for P2-M is comparable to the correlation skill obtained for the ECMWF system (version Cy32r3), where correlation skill greater than 0.5 is obtained out to 23 days (Vitart and Molteni 2011). The correlations from P2-M are significantly different from P2-S (but not from P1.5) over days 11–30 of the forecast [5% significance level, two tailed, Fisher’s Z transformation; \( n = 648 \) (i.e., 12 start months, 27 yr, and two independent components comprising the MJO index)].

The SAM is an important mode of variability for high and middle latitudes and is characterized by shifts in the strength of the zonal flow between about 55°–60°S and 35°–40°S (e.g., Gong and Wang 1999; Thompson and Wallace 2000). To assess the skill in predicting the SAM, a daily SAM index is calculated for the observations by projecting observed daily mean sea level pressure (MSLP) anomalies (NCEP–NCAR reanalysis; Kalnay et al. 1996) onto the leading EOF of observed zonally averaged monthly mean MSLPs between 25° and 75°S. The POAMA SAM index is calculated by projecting the daily mean predicted MSLP anomalies from each POAMA ensemble member onto the observed EOF [see Marshall et al. (2011c) for details]. Figure 12 shows the ensemble mean RMSE and correlation, and the ensemble spread for the three forecast systems. For a climatological forecast of the SAM index, the RMSE = 1. P2-M reaches this threshold after about 17 days, whereas P1.5 reaches it after 14 days (Fig. 12a). As seen for the MJO index, P2-S overshoots this climatological threshold due to the lack of ensemble spread (Fig. 12a). The P2-M system has significantly more spread than the P2-S system and more spread than P1.5 for the first 15 days of the forecast. Interestingly, the spread in P2-M actually exceeds the error (overdispersive) for the first 5 days of the forecast. Correlation skills above 0.5 are achieved out to 12 days for P2-M and out to 10 days in P1.5 (Fig. 12b). The new P2-M model therefore has a slightly improved ability to predict the SAM index. The correlations from P2-M are significantly different from P2-S (but not from P1.5) over days 14–18 of the forecast [5% significance level, two tailed, Fisher’s Z transformation; \( n = 324 \) (i.e., 12 start months and 27 yr)].

4. Contribution to improvement: Ensemble generation strategy, multimodel formulation, or ensemble size?

The section above has shown that the P2-M system is more skillful and reliable than both the P1.5 and P2-S
systems for intraseasonal and even the first season forecasts. The major difference between P2-M and P2-S is the new ensemble generation strategy, which provides perturbations to both the atmosphere and ocean at the initial time (P2-M) compared to only perturbations to the ocean (P2-S). The other minor difference is that P2-M has 33 ensemble members and P2-S has 30 members (with 11 and 10 unique initial conditions, respectively). It seems clear, therefore, that the improvements in P2-M over P2-S can be ascribed to the new ensemble generation technique. However, the reason for the improvement of P2-M over P1.5 is not yet clear: it could be due to the coupled-breeding ensemble generation technique of P2-M compared to the lagged-average ensemble method of P1.5; it could be due to the larger ensemble size of P2-M (33 members) compared to P1.5 (10 members) or it could be due to the benefits of the multimodel configuration of P2-M compared to the single model of P1.5.

Figure 13 shows the ROC area, as well as the resolution (RES) and reliability (REL) components of the Brier score, for rainfall over Australia above the upper tercile. The Brier score can be decomposed into three components, two of which measure forecast resolution and reliability (Murphy 1973). The third term is a function of the climatological frequency of the occurrence of the event and is not related to forecast quality. A forecast system has good resolution and a high value for the RES component if the forecast probabilities can discern subsample forecast periods with different relative frequencies of the event [i.e., frequencies that differ from the climatological base rate; Wilks (2006)]. In terms of the attributes diagram, RES measures the deviation of the curve from the horizontal line that intersects the y axis at the climatologically observed frequency. If a system had no resolution, the reliability curve would lie on this horizontal "no resolution" line (Wilks 2006). In contrast, the REL component measures the conditional bias of the probabilistic forecasts (i.e., the correspondence between the forecast probability and the relative observed frequency). In terms of the attributes diagram, REL measures the deviation of the curve from the diagonal 1:1 line, such that perfect reliability will give \( \text{REL} = 0 \).

The conclusions from the ROC area and RES results in Fig. 13 are very similar to each other [Toth et al. (2003) note that the resolution component of the Brier score and the ROC area often provide very similar qualitative information of the forecasting system]. In fortnight 1 for all seasons, the P2-M system has the highest discrimination (ROC area) and resolution (RES) compared to the other two systems, whereas the differences between the systems are smaller in fortnight 2 (consistent with Fig. 4). The multimodel configuration, which also implies a larger ensemble size, has little or no impact on the forecast resolution and discrimination, shown by negligible differences between the constituent models and the relevant multimodel for both P2-M and P2-S, respectively (Figs. 13a, 13b, 13d, and 13e). We note that ensemble size alone does not normally have a large impact on forecast resolution (Doblas-Reyes et al. 2008).

The biggest difference between the three systems and within each system is related to forecast reliability. It is clear that P2-M is the most reliable system for fortnights 1 and 2 in all seasons, but there are other important inferences that can be made from Figs. 13c and 13f. First, the benefit of the multimodel and larger ensemble size (compared to the constituent models) is quite small for
both P2-M and P2-S in the first fortnight. This is also
evident in the previously shown plot of RMSE and en-
semble spread (Fig. 3b; P2-S and 3c P2-M), where the
gray lines, giving the result from one of the constituent
10-member ensemble models, can be contrasted with
the colored lines showing the larger MME ensemble.

<table>
<thead>
<tr>
<th>Month</th>
<th>DJF</th>
<th>MAM</th>
<th>JJA</th>
<th>SON</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROC Area</td>
<td>1.0</td>
<td>0.6</td>
<td>0.4</td>
<td>0.2</td>
<td>0.0</td>
</tr>
<tr>
<td>RES</td>
<td>0.10</td>
<td>0.08</td>
<td>0.06</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>REL</td>
<td>0.10</td>
<td>0.08</td>
<td>0.06</td>
<td>0.04</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Figure 13 does, however, suggest that in fort-
night 2 the larger MME has a more significant positive
contribution to forecast reliability in relation to the
constituent models compared to fortnight 1. As men-
tioned in the previous section, on the seasonal time scale
the benefit of the larger MME on forecast reliability is
clear (Wang et al. 2011). Seasonal forecasts (3-month
average with a 1-month lead time) of Australian rain-
fall above the upper tercile from the large multimodel
are reliable, whereas those from the constituent models
are not (Wang et al. 2011). Wang et al. (2011) found
that the major contribution to the improved reliability
of the P2-S multimodel compared to the component
models is from the use of three model versions rather
than from an increased ensemble size. For instance,
ENSO behaves differently in each of the component
models of POAMA-2 and this will affect the simulated
teleconnections to regional rainfall and temperature
(Wang et al. 2011).

The reason for the modest contribution of the larger
ensembles of P2-S and P2-M to the forecast spread and
reliability on intraseasonal time scales is because their
component models are initialized using the same set of
initial conditions; that is, for P2-M (P2-S) there are only
11 (10) unique initial conditions available for initializa-
tion. The impact of this is illustrated in an example of an
MJO forecast using P2-M (Fig. 14). In general, there is
less separation of the forecast plumes when grouped
by ensemble member (Fig. 14a) than when grouped by
constituent model (Fig. 14b) (cf. forecasts with the same
color in Fig. 14). For the former, there are distinct groups
of three forecast plumes that indicate ensemble members
from the three constituent models have been initialized
with the same initial conditions (Fig. 14a). This clearly
signifies the need for using different initial conditions
for each of the three models, to better simulate the en-
semble spread at these time scales. For the next version

![Figure 13](image-url)
of POAMA we plan to use a unique initial condition for each member.

The component models P2c-M and P2c-S have exactly the same model formulation as P1.5 and their ensembles are of comparable size. Comparing these three models therefore highlights differences due to ocean initial conditions and the ensemble generation strategies. For both fortnights and all seasons, P1.5 is more reliable than P2c-S, but less reliable than P2c-M (Figs. 13c and 13f). P2c-M and P2c-S both use ocean initial conditions from the PEODAS assimilation, whereas P1.5 uses ocean initial conditions from the optimum interpolation assimilation system. The differences in reliability and skill (the latter is discussed in the next paragraph) between the three models do not, therefore, appear to be related to the quality of the ocean initial conditions [which are superior using PEODAS; Yin et al. (2011); Zhao et al. (2013b)]. This does not imply that the ocean initial conditions have no impact or are not important. In fact, P2-S shows better skill at predicting seasonal sea surface temperature variations associated with ENSO compared to P1.5, directly as a result of the improved ocean initial conditions (Zhao et al. 2013a, manuscript submitted to Climate Dyn.). However, on intraseasonal time scales, the main source of the difference is related to the ensemble generation strategy. P2c-S is the least reliable because, as discussed previously, perturbing only the ocean initial conditions and not the atmosphere does not produce enough ensemble atmospheric spread on intraseasonal time scales. P2c-M has improved reliability over P1.5 due to the different ensemble generation technique, namely a burst ensemble with perturbed atmosphere and ocean initial conditions compared to a time-lagged ensemble (with no ocean perturbations). The improved ensemble generation strategy for P2-M provides a clear benefit in terms of greatly increased reliability compared to P2-S and P1.5.

This conclusion is also supported by the Brier skill score calculation (Fig. 15), which shows the percentage improvement in skill compared to a climatological forecast for forecasting above the upper tercile for Tmax in the first and second fortnights for P1.5, P2c-S, and P2c-M (i.e., all three have exactly the same model configuration and ensemble size). For both fortnights there is a clear improvement in skill in P2c-M compared to P1.5 and P2c-S (the same is true for Tmin and rainfall; not shown). P2c-M shows 20%–35% increases in skill compared to a climatological forecast in the first fortnight and 10%–15% improvements in the second fortnight, with improvements focused over western and southeastern Australia (Fig. 15).

5. Conclusions

The production of skillful and reliable probabilistic forecasts is a desirable attribute of any forecasting system. The new POAMA-2 multiweek forecast system (P2-M) demonstrates greatly improved reliability of forecasts of Australian climate for lead times out to one season, in comparison to the seasonal configuration of
FIG. 15. Debiased BSS (reference score is climatology) for 2-week mean maximum temperature above the upper tercile for (left) fortnight 1 (days 1–14) and (right) fortnight 2 (days 15–18) using all forecast start months for (a),(b) P1.5, (c),(d) P2c-S, and (e),(f) P2c-M. The score reflects the percentage (i.e., multiply the score by 100) improvement in skill compared to a climatological forecast. Grid boxes showing skill better than a climatological forecast by more than 10% are shaded. Areas of negative skill are shown by the dashed contours. The contour interval is 0.05 (i.e., 5%). These three single model systems differ in their ensemble generation strategy, but have the same model configurations and ensemble size (10 members).
POAMA-2 (P2-S) and the older POAMA-1.5 (P1.5) version. This improvement can be attributed to the new ensemble generation scheme used for characterizing uncertainty in initial conditions. This scheme, applied to the forecasts in burst ensemble mode (all ensemble members initialized at the same time), generates perturbed atmosphere and ocean initial conditions using a coupled-breeding approach. In contrast, P2-S uses a burst ensemble created from perturbed ocean initial conditions, but no perturbations are applied to the atmosphere; and P1.5 uses time-lagged atmospheric initial conditions to generate the ensemble, but the ocean initial condition is the same in all members. The coupled-breeding scheme results in increased ensemble spread in the first month of the forecasts compared to either the P2-S approach or the lagged-ensemble approach of P1.5, and this reduces the underdispersive nature of the forecasts. Although the major impact of the new ensemble generation scheme is to improve the reliability of forecasts, the P2-M system also exhibits improved forecast discrimination and skill in fortnights 1 and 2 compared to P2-S and P1.5, as shown by the ROC area and Brier skill score maps. It also demonstrates increased skill in predicting important drivers of intraseasonal climate, namely the MJO and the SAM. The benefits of the coupled ensemble generation approach are not restricted to the intraseasonal time scale. Forecasts of zero-lead seasonal mean rainfall and temperature over Australia are more reliable and skillful in P2-M compared to P2-S. For this reason, the BoM’s operational seasonal forecasts from POAMA are now based on the P2-M system configuration, thus having one system producing a seamless set of products from weeks to seasons.

The results of our study demonstrate the possible shortcomings of using lagged atmospheric initial conditions as a method of generating perturbed initial conditions. Lagged-atmospheric initial conditions, which is a common and simple approach to generating perturbations (e.g., Jones et al. 2011; Pegion and Sardeshmukh 2011; Liu et al. 2012), can result in underdispersive forecasts (assuming no additional method of perturbing the forecasts), even though the magnitudes of the initial perturbations are comparable or larger than the analysis uncertainty. Presumably, the increased spread and faster growth in spread from the breeding method are because the method is more effective at emphasizing the fast-growing errors. The breeding method is designed to perturb those modes that grow the fastest during the analysis cycle leading up to the initial time and are thus the modes most likely to dominate the analysis errors (e.g., see also Toth and Kalnay 1993). Our study has shown, through a comparison of the ensemble spread with the error of the ensemble mean, that the breeding method has a better-tuned ensemble compared to the lagged ensemble. This does not preclude using a combined approach to generate the final ensemble [i.e., combining successive times of burst (bred) ensembles to create a larger ensemble]. The substantial improvement in spread/dispersion from using perturbations produced by coupled breeding does not, however, indicate that these perturbations are optimal in the sense that the full range of possible outcomes is adequately sampled given the limited ensemble size. Ongoing experimentation is aimed at optimizing the method of generating perturbations using coupled breeding, but the gains made to date already justify the approach. We also note that our comparison between the breeding method (i.e., P2c-M) and the lagged method (P1.5) is not a clean comparison, since the ocean assimilation systems are different in the two systems (note that P2c-M and P1.5 are identical models with comparable ensemble sizes). We are currently running additional experiments that isolate the impact of the ensemble generation method, including answering other questions such as the importance of having flow-dependent perturbations.

Unlike for seasonal time scales, on intraseasonal time scales there appears to be little benefit from the POAMA-2 multimodel approach. This is because each component model version uses the same set of perturbed initial conditions. As such, the benefit of using different model formulations is only evident at much longer lead times for which ensemble spread develops as a result of model differences (Wang et al. 2011). Future development of POAMA will include using different perturbed initial conditions for each model version, to better simulate the ensemble spread at these time scales.

Like most intraseasonal–seasonal forecast systems, P2-M still has separate data assimilation schemes for the ocean, land, and atmospheric components of the global coupled model. Separate assimilation can lead to initialization shock in the forecasts, as the initial coupled model state is not in dynamical balance. In contrast, a fully coupled data assimilation scheme would likely reduce the shock and improve forecasts. The ensemble generation and initialization strategy developed for P2-M provides a solution somewhat closer to that which would be obtained with fully coupled assimilation and is closely aligned with the long-term goal of fully coupled data assimilation with POAMA. Since the breeding scheme is performed with the coupled model, the perturbations that are used to initialize the forecasts are in dynamical balance. This new strategy represents a significant milestone in the development of the POAMA forecast system. How best to initialize the coupled system for successful intraseasonal prediction is still largely an unanswered question. This
work contributes to advancing our knowledge on this issue.

Acknowledgments. This work was supported by the Managing Climate Variability Program of the Grains Research and Development Corporation. Mei Zhao, Wasyl Drosdowsky, and two anonymous reviewers are thanked for their useful comments in the preparation of this manuscript.

REFERENCES


