Tellus (2005), 57A, 320–339
Printed in UK. All rights reserved

TELLUS

A performance comparison of coupled and uncoupled versions of the Met Office seasonal prediction general circulation model

By R. J. GRAHAM[1]*, M. GORDON[1], P. J. MCLEAN[1], S. INESON[1], M. R. HUDDLESTON[1], M. K. DAVEY[1,2], A. BROOKSHAW[1] and R. T. H. BARNES[1], [1]*Met Office, FitzRoy Road, Exeter, EX1 3PB, UK;* [2]*University College London, UK*

(Manuscript received 5 April 2004; in final form 15 November 2004)

ABSTRACT

We compare the performance of the Met Office's ocean–atmosphere coupled general circulation model (CGCM) seasonal prediction system with that of an atmosphere-only system (AGCM). The CGCM and AGCM systems share the same atmospheric component and the performance comparison therefore provides insight into the skill benefits available from coupling atmosphere and ocean models. In this study, the AGCM is forced with predicted sea surface temperature (SST) based on persistence of prior observed SST anomalies. The analysis uses 43-yr, nine-member ensemble hindcast data sets generated with both systems as part of the European Union project DEMETER. Results are focused on global and regional comparisons of long-term skill for probabilistic prediction of 2-m temperature in the upper tercile, and on selected case studies for the tropics and Europe.

Performance assessments using relative operating characteristic scores, Brier skill scores and the resolution and reliability terms of the Brier score decomposition are contrasted. The largest CGCM benefits are found in tropical regions, where benefits to both resolution (essentially 'event detection') and to reliability (essentially 'calibration' of the forecast probabilities) are demonstrated. Improvements to reliability are found to be substantially greater than improvements to resolution. Regional assessments show benefits, as expected, in the tropical east Pacific, from improved prediction of SST variability associated with the El Niño Southern Oscillation (ENSO). However, substantial benefits are also seen throughout the tropical belt in seasons associated with the peak and decay of ENSO activity. Such benefits appear associated with representation of lagged teleconnection responses to ENSO in the tropical Atlantic and Indian Oceans.

In the extratropics, CGCM improvements to reliability are also substantial, although benefits to resolution (assessed over large regions) appear negligible. Two classes of benefit are described. First, advantages from improved ENSO predictions appear to benefit skill in the North Pacific and North American regions, through teleconnection responses. Secondly, there is evidence of benefits from representation of coupled processes over the North Atlantic. In particular, CGCM skill benefits for prediction of spring season temperature in the European region appear to derive, in part, from coupled model representation of linkage between a well-documented tripole pattern in North Atlantic SST anomalies and the North Atlantic oscillation. This result provides encouraging evidence that use of CGCMs offers prospects for improving seasonal prediction in the extratropics through representation of coupled ocean–atmosphere processes in extratropical ocean basins, as well as through indirect impacts from improved prediction of ENSO and associated teleconnections.

1. Introduction

In recent years, a number of meteorological centres have begun routine production of global seasonal forecasts using long integrations of general circulation models (GCMs). Two main GCM configurations are used: coupled ocean–atmosphere models (CGCMs; e.g. Stockdale et al., 1998) and 'two-tier' systems in which an atmosphere-only model (AGCM) is forced with an independent prediction of the sea surface temperature (SST) evolution. Predicted SST in two-tier systems is often based on simple persistence of observed SST anomalies, but may also employ more sophisticated statistical methods for tropical SST prediction (e.g. Barnston et al., 2003). In this paper we compare the performance, for global-scale seasonal prediction, of a CGCM and an AGCM forced with a persisted SST anomaly

*Corresponding author.
e-mail: richard.graham@metoffice.gov.uk

(PSSTA) based prediction of SST. The comparison is made with reference to long-term skill scores and selected case studies obtained from retrospective forecasts with both systems over the 43-yr period (1959–2001). The retrospective forecasts (or hindcasts) were generated as part of the European Union (EU) project DEMETER (Palmer et al., 2004). We restrict assessments to predictions of 2-m temperature, because the greater predictability for temperature (relative to precipitation, the other main variable of interest) provides a clearer signal for the case study investigations.

The CGCM used is the current Met Office global real-time seasonal prediction model (known as GloSea). The AGCM used is the Met Office HadAM3 model, which is the atmospheric component of GloSea. HadAM3+PSSTA (hereafter referred to simply as HadAM3) was formerly used for real-time prediction prior to being replaced by GloSea in March 2003. Because the same atmospheric component is used in both systems, performance differences will derive only from the impact of ocean–atmosphere coupling in the GloSea system. The impact from coupling enters both through the forecast evolution and also through the ensemble perturbation strategies employed, because generation of perturbations in the GloSea system makes use of the coupling, while the strategy for HadAM3 must necessarily use other methods. Thus, the magnitude of the performance benefits due to coupling will depend to a degree on the limitations of the HadAM3 perturbation strategy. However, because the perturbation strategies used in both systems are similar to those used in other operational centres, the performance comparison will provide insights into the benefits of CGCM systems over two-tier AGCM systems in general. We note also that PSSTA-based systems (such as HadAM3) will potentially provide a less skilful reference than two-tier systems forced with SST from empirical prediction systems, or from independent dynamical ocean predictions (e.g. Barnston et al., 2003; we shall refer to such systems as FSST systems). However, although ocean to atmosphere forcing is potentially improved, on average, in FSST systems relative to PSSTA systems, atmosphere to ocean forcing, important for example in promulgating remote responses to El Niño Southern Oscillation (ENSO) SST anomalies, is absent where statistical SST prediction is used. Thus, CGCM benefits documented here relative to PSSTA systems are also likely to be apparent relative to many FSST systems, although smaller in magnitude. Two-tier seasonal prediction systems are in wide operational use, and because CGCMs are significantly more costly than two-tier systems, both to develop and to operate, assessments of the performance benefits, the subject of this paper, are needed to weigh against the practical costs.

Potential CGCM benefits to atmospheric seasonal prediction lie through improved prediction of the SST field, which plays a fundamental role in modulating seasonal atmospheric anomalies (Palmer and Anderson, 1994). Large-scale anomalies in SST develop largely as ocean surface manifestations of slowly evolving (over seasonal time-scales) three-dimensional (3-D) ocean circulations that are driven by fluxes of momentum, heat and fresh water from the atmosphere. Thus, by modelling the full 3-D ocean and addressing ocean–atmosphere interactions, CGCMs offer opportunities for improving and extending the range of SST prediction, and therefore atmospheric seasonal prediction, which are unavailable to the two-tier approach. To date, development of CGCMs has been largely motivated by a desire to improve prediction of tropical Pacific SST warming and cooling events associated with the El Niño/La Niña cycle (or ENSO), such events having well-documented influences on global-scale seasonal climate anomalies (e.g. Ropelewski and Halpert, 1987). Some notable cases of current CGCM successes in predicting such events have been reported (e.g. Stockdale et al., 1998; Palmer et al., 2004). In contrast, PSSTA SST predictions will inevitably fail in the tropical Pacific when El Niño or La Niña SST anomalies develop or decline significantly over the forecast range. However, despite this (important) shortcoming, PSSTA methods frequently provide competitive SST predictions in other ocean basins, because of the typically slow observed evolution time-scales. PSSTA can also be competitive in the tropical Pacific over some parts of the ENSO cycle, for example, when extrapolating an existing, slowly developing, ENSO SST anomaly. In such circumstances, the main advantage of CGCMs is their potential to capture the developing amplitude of the event, through modelling of ocean–atmosphere feedback mechanisms, and, through the atmospheric 'bridge', SST responses to ENSO in other ocean basins and their associated atmospheric responses.

Coupled model performance for prediction of tropical Atlantic and Indian Ocean SST anomalies has received less attention than studies of ENSO predictability. However, (Davey et al., 2002, hereafter D02) have demonstrated the ability of a number of current coupled models to reproduce observed lagged correlations of tropical Atlantic and Indian Ocean SST anomalies with ENSO SST anomalies in the tropical Pacific. There is little documentation on current CGCM performance in predicting extratropical SST anomalies, although such anomalies may impact on extratropical seasonal climate. Recent investigations of the role of extratropical oceans in seasonal predictability include those of (Rodwell and Folland, 2002, 2003, hereafter RF03) and Rodwell et al. (1999) who identify a tripole anomaly pattern in North Atlantic SST associated with northern winter and spring season variability in the North Atlantic Oscillation (NAO) index, a climate index correlated with seasonal variability of temperature and precipitation over Europe (Hurrell, 1995). Sutton and Hodson (2003) find evidence of an influence from both North Atlantic and tropical Pacific ENSO SST variability on seasonal climate variability over the North Atlantic/European region.

Section 2 provides a summary of the DEMETER experimental configuration and key features of the GloSea and HadAM3 systems. A brief discussion of the systematic errors in the two systems, and comparisons of prediction skill for Niño3.4 region SST anomalies are provided in Sections 3 and 4, respectively. The frequency of significant differences in the ensemble

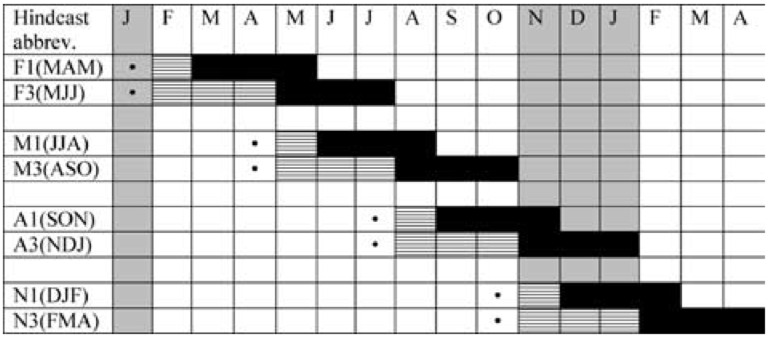

Fig 1. Position of the DEMETER hindcast periods in the annual cycle. The lead time (hatching) and three-month periods verified (black shading) for the February-, May-, August- and November-start hindcasts are shown. Abbreviations in the left-hand column are as defined in the text. For HadAM3, SST prediction is based on persistence of SST anomalies observed in the months indicated with dots. For reference, the three-month period in which tropical Pacific ENSO SST anomalies typically reach peak magnitude (NDJ) is shown (light shading).

probability distribution functions (pdfs) of 2-m temperature anomaly, obtained with each system, is investigated in Section 5. Section 6 provides a comparison of long-term prediction skill for 2-m temperature in the upper tercile, using standard performance measures. In Section 7 we present two case study investigations exploring instances of different behaviour between GloSea and HadAM3 in the tropics and Europe. A summary and discussion of results is given in Section 8.

2. Experimental set-up

The GloSea and HadAM3 seasonal hindcast data sets used in this study are those generated for the 43-yr period 1959–2001 as part of the DEMETER project. For each year, the models were run in nine-member ensembles to a range of six months from four start dates: 00 GMT on the first day of February, May, August and November (Fig. 1).

2.1. *GloSea CGCM experiments*

The GloSea model is based on the Met Office coupled climate model HadCM3 (Gordon et al., 2000). It uses a stretched north–south ocean grid, in which a resolution of 1.25° in both the meridional and zonal directions improves to 0.28° in the meridional direction in the tropics, and has 40 vertical levels. No flux correction is employed, and an interactive sea-ice model is incorporated. The atmospheric component of GloSea is HadAM3 (Pope et al., 2000), the same model used for the two-tier experiments. HadAM3 has a horizontal resolution of 2.5° latitude by 3.75° longitude and 19 vertical levels. GloSea uses a coastal tiling scheme that allows the ocean model to have a coastline determined by the ocean grid, rather than by the lower resolution atmosphere grid, giving improved representation of land/sea geography.

Initial ocean conditions for the GloSea hindcasts were obtained by forcing the ocean component with momentum, heat and fresh water fluxes from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA40 reanalysis, after first spinning up from climatological values of temperature and salinity (Levitus and Boyer, 1994; Levitus et al., 1994) using ERA40 flux climatologies. A slow time-scale relaxation of temperature and salinity to climatology is applied at all model levels. From 1986 onwards, the FOAM ocean data assimilation scheme is used to assimilate subsurface temperature data. The FOAM scheme (described for short-term ocean prediction by Bell et al., 2000) uses a form of optimal interpolation. An equatorial pressure gradient bias-correction scheme (Bell et al., 2004; Huddleston et al., 2004) is used to address unrealistic deep overturning circulations that can occur near the equator when only thermal data are assimilated into an ocean model.

To generate perturbed initial conditions for the ensemble, we use the method described in Palmer et al. (2004). The method is the same as used for real-time forecasting, but adapted here for the nine-member ensembles (40 members are used in real-time predictions). Briefly, sets of wind stress and SST perturbations, designed to represent observed uncertainties in these parameters, were predefined using historical differences between quasi-independent analysis data sets. A series of wind stress perturbations, sampled from the predefined set, was applied with positive and negative signs to the ERA40 momentum fluxes used in the ocean assimilation, resulting in one unperturbed and two perturbed assimilation sequences. At each hindcast start time, instantaneous SST perturbations are added to the three assimilation sequences. This is achieved by selecting (at random) four SST perturbations from the predefined set and applying them with positive and negative signs in such a way as to generate eight perturbed initial ocean states and one unperturbed control state. The SST perturbations are added to the temperature field in the top 40 m of the ocean, with a ramp to zero at 40 m. Thus, perturbations are applied to the ocean initial state only, with unperturbed ERA40 analyses being used to initialize the model atmosphere and land components.

2.2. *Two-tier HadAM3 AGCM experiments*

In the two-tier experiments, HadAM3 is forced with a PSSTA SST prediction scheme. The PSSTA scheme includes treatment of sea ice and is based on persistence of the one-month SST anomaly (derived from HadISST1; Rayner et al., 2003) in the month preceding the integration start date. The SST anomalies are added to the monthly HadISST1 SST climatology for the period of integration, interpolated in time and updated in the model at 24-h intervals. In the open ocean, SST anomalies are kept constant until month four and then relaxed to zero by the end of month six. Ice is assumed to be present where predicted SST falls below −1.8°C; thus, a positive (negative) initial SST anomaly can delay (hasten) ice formation.

In contrast to the GloSea experiments, perturbations for the two-tier system are derived using atmospheric perturbations only, the SST forcing being identical for all ensemble members. A lagged approach is used in which the ERA40 analyses at the hindcast start time and eight preceding analyses (at six-hourly intervals) form the initial atmosphere and land conditions for the nine members.

The GloSea and HadAM3 ensemble strategies used reflect operational practice with the two systems. Closer similarity in the perturbation methods might have been achieved by forcing HadAM3 with SST anomalies derived from the nine GloSea initial ocean states. In practice, however, the initial SST perturbations generated in the GloSea system are small in comparison with the interannual SST variations that largely determine the ensemble evolution. Consequently, use of SST-only or (as employed here) atmosphere-only initial perturbations is unlikely to influence greatly the HadAM3 results.

2.3. *Verification methods and forecast bias correction*

The performance of GloSea and HadAM3 hindcasts for Niño3.4 region SST anomaly and for 2-m temperature over tropical and extratropical regions are compared, with emphasis on the latter. Verification is performed for three-month-mean values corresponding to months 2–4 (one-month lead) and months 4–6 (three-month lead) of the integrations. To ease association of each three-month valid period with the hindcast start month and lead time, we use the convention that, for example, F3(MJJ) refers to the February-start hindcasts at three-month lead, valid for the period May–June–July. The start and range of the hindcast sets and the three-month periods evaluated are shown in Fig. 1. A local linear bias correction is applied (a posteriori) by expressing the forecasts relative to the model climatology for the appropriate hindcast/lead time, as defined from 378 realizations (42 yr × nine members). In all cases, the cross-validation technique is used, in which the (42-yr) reference climatology for each hindcast is constructed without the year of the hindcast itself. Cross-validation simulates typical real-time operational practice of using a fixed period for the reference model climatology, which does not extend to the current forecast year, and therefore provides a more faithful representation of real-time skill.

ERA40 is used to verify both the SST and 2-m temperature predictions. For assessment of deterministic skill for Niño3.4 SST anomalies, GloSea ensemble-mean and PSSTA SST predictions are averaged over all model grid points in the Niño3.4 region (5°S–5°N, 170°W–120°W) and correlated with the corresponding ERA40 average. For 2-m temperature, probabilistic predictions of equiprobable tercile categories (i.e. below-normal, near-normal and above-normal) have been generated. For brevity, we present skill assessments for the upper tercile only, because skill was found to be similar for the upper and lower terciles (at least when aggregated over large geographical areas), and because skill for the middle tercile is generally substantially lower than for the outer two, and therefore of less interest in terms of useability. Relatively low skill for the middle tercile has been reported by numerous other authors (e.g. Van den Dool and Toth, 1991).

The assessment measures used for probabilistic forecasts are the area under the relative operating characteristic curve or ROC score (Stanski et al., 1989), reliability and sharpness diagrams, the reliability and resolution components of the Brier score (BS), and the Brier skill score (BSS). The reader is referred to Toth et al. (2003) and Wilks (1995) for the numerical formulation and a fuller discussion of these performance measures. The boundaries of the forecast tercile categories are defined for each model grid point by interpolation in the model climatology (using cross-validation). The boundaries defining the observed categories are defined in the same way using the ERA40 reanalysis from the verification data sets. Forecast probabilities for the tercile categories, derived from the ensemble, were stratified into 10 probability bins (p_n) defined by $0\% \leq p_1 < 10\%$, $10\% \leq p_2 < 20\%$..., $90\% \leq p_{10} \leq 100\%$. A global comparison of skill is provided with emphasis on three regions: the tropics (30°N to 30°S), the northern extratropics (30°N to 90°N) and Europe (12.5°W to 42.5°E; 35°N to 75°N).

3. GloSea and HadAM3 SST and 2-m temperature biases

Model biases may impact on predictability in complex ways, generating errors that are not addressed by the linear local bias correction applied. In this section we provide some general comments on SST and 2-m temperature biases in the GloSea and HadAM3 systems. Predictions of temporal variations in SST anomaly with CGCMs are typically associated with systematic biases (see, for example, D02). Such biases will contaminate forcing of the atmosphere by the ocean, and may detract from CGCM benefits for prediction of atmospheric variability. In contrast to CGCMs, PSSTA methods have zero biases (because of ‘anchoring’ to the observed climatology) but do not, by definition, attempt to predict variations in SST anomaly over the forecast range except (in this study) for a fixed rate of anomaly decay after month 4 (Section 2). Thus, when comparing the performance of GloSea and HadAM3 for atmospheric predictability, we expect a trade-off between the beneficial impacts of reduced GloSea error in prediction of SST anomaly variability, and negative impacts from increased GloSea systematic errors.

A full description of GloSea biases is beyond the scope of this paper. In general, SST biases were found to be comparable in magnitude and distribution to other participating DEMETER CGCMs (biases for a wide range of variables may be viewed on the DEMETER website http://www.ecmwf.int/research /demeter/d/charts/verification). Over land, GloSea and HadAM3 biases in 2-m temperature are generally very similar in location and magnitude in all sets of hindcasts, as is to be expected from use of the same atmosphere and land-surface components.

4. Prediction of Niño3.4 SST anomalies

Over the oceans, prediction skill for 2-m temperature, assessed in Section 6, will closely reflect skill for SST. Thus, for brevity we explicitly assess GloSea and HadAM3 SST predictions only in terms of correlation skill for Niño3.4 SST anomalies. Skill is best with both systems in the August and November hindcasts (Table 1; correlations at both lead times exceed or equal 0.9 with GloSea and 0.8 with PSSTA), consistent with the fact that in years with active ENSO events tropical Pacific SST anomalies will typically be present at initialization and throughout much of the predicted range. However, skill differentials between the two systems are lowest in these hindcasts, indicating that damped persistence of July and October SST anomalies can provide competitive performance. Skill with each system is similar at both lead times in the August hindcasts, but drops more markedly for PSSTA at three-month lead in the November hindcasts [N3(FMA)], consistent with GloSea superiority in prediction of decaying ENSO SST anomalies. The largest advantages with GloSea are seen in the February and May hindcasts, with notable benefits at both lead times in the May hindcasts. ENSO SST anomalies are typically in an early stage of development around the May initialization time, and the large skill differential implies benefits from coupled model potential to develop incipient ENSO SST anomalies into the maturing phase. Low skill from PSSTA in the May hindcasts is not surprising, because persistence of April SST anomalies will generally be a poor strategy in years with active ENSO events.

Table 1. Correlation of GloSea ensemble-mean (G) and PSSTA (P) three-month Niño3.4 SST anomalies with corresponding observed (ERA40) anomalies for the period 1959–2001, at one-month and three-month leads

	Lead time			
	One-month		Three-month	
Hindcast/period	G	P	G	P
Feb: F1(MAM), F3(MJJ)	0.89	0.81	0.77	0.28
May: M1(JJA), M3(ASO)	0.76	0.47	0.76	0.33
Aug: A1(SON), A3(NDJ)	0.90	0.91	0.90	0.88
Nov: N1(DJF), N3(FMA)	0.95	0.94	0.92	0.83
All	0.89	0.78	0.85	0.54

5. Frequency of differences in GloSea and HadAM3 ensemble distributions

Before presenting skill score results for predictions of the upper tercile, we first examine the frequency of differences in the pdfs for 2-m temperature anomalies, as represented by the nine-member ensembles of each system. Notable differences in skill between the two systems are only likely at locations where the two pdfs vary significantly in a sufficient fraction of the 43-yr sample. The number of years for which the two pdfs are found different at the 95% level (at any given grid point) is shown for the M3(ASO) period in Fig. 2 (the Wilcoxon–Mann–Whitney test is used; see, for example, Wilks, 1995). Results for the other hindcasts were found to be similar. For much of the tropical ocean and land regions, significant differences occur in over half the sample (i.e. in ~20 yr or more) with peaks exceeding two-thirds of the sample most widespread in the tropical Pacific. The high frequency of differences implies ample opportunity for differences in prediction skill between the two systems. In contrast, significant differences in the pdfs of the two systems are less frequent in the extratropics, occurring typically in 10–20 yr over oceans and generally in less than 5 yr (independently at each grid point) over continents. The differential in frequencies between the extratropical ocean and land regions suggests that in many cases pdf differences over oceans are probably associated with differing spread characteristics rather than with displacements in the mode of the distributions. Larger ensemble spread in GloSea over the oceans can be expected because, unlike in HadAM3, ocean perturbations are applied initially and are free to develop through the ocean–atmosphere coupling. Such differences appear rapidly attenuated over land by virtue of the common land and atmospheric components of the two systems. The relatively few occasions when significant differences are present over land are more likely to correspond with instances when major circulation differences have developed between the two systems, resulting in significant differences in the location of the mode of the two pdfs. In Section 7 we make use of time series of pdf differences to identify examples for case study.

6. Performance comparison for probabilistic prediction of 2-m temperature in the upper tercile

Different skill scores will in general assess different aspects of forecast performance, and it is therefore advisable to base evaluations on a number of different measures. In this section we present a comparison of GloSea and HadAM3 performance using ROC scores, reliability diagrams and BSS. In addition, a comparison of reliability and resolution characteristics is provided and used to interpret contrasting results obtained with ROC scores and BSS. As discussed by Toth et al. (2003), resolution and reliability are the two attributes which together determine the usefulness of probabilistic forecast systems. Briefly stated,

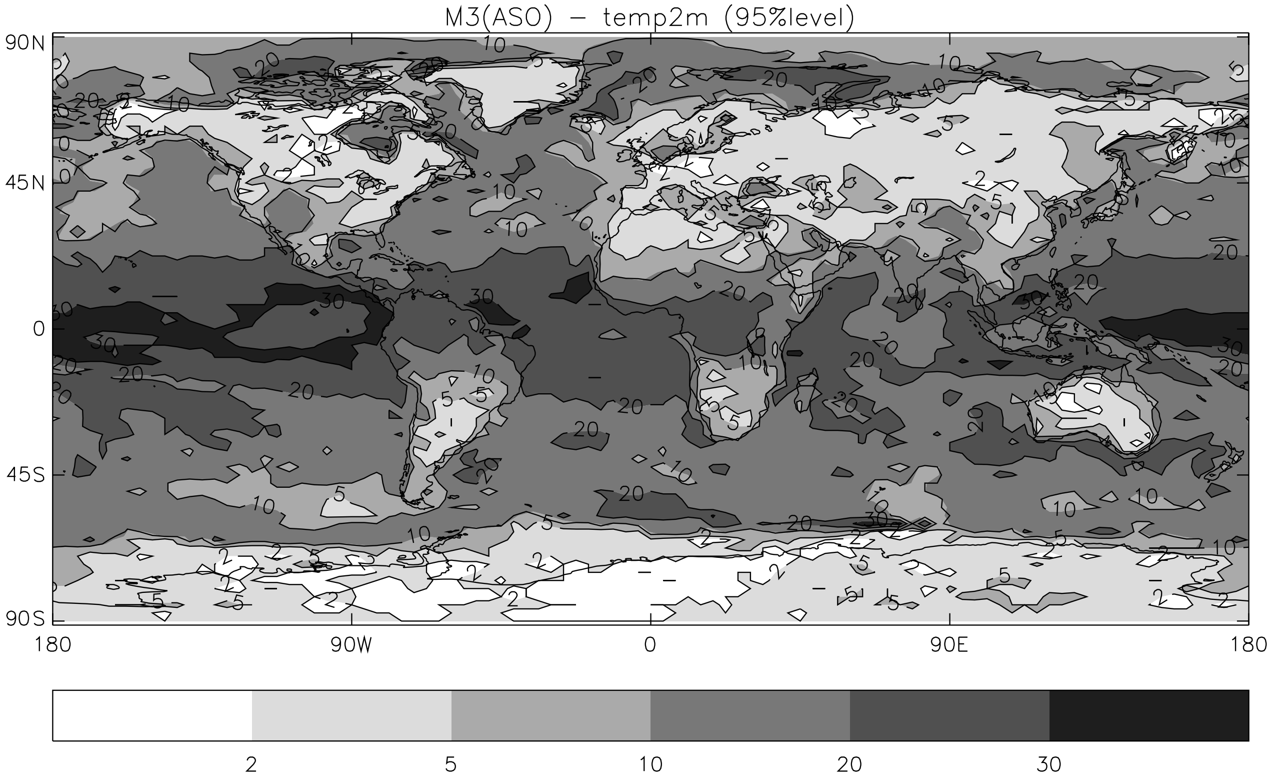

Fig 2. Number of years (out of 43, 1959–2001) when GloSea and HadAM3 nine-member ensemble distributions for M3(ASO) 2-m temperature anomalies were found significantly different at the 95% level with a Wilcoxon–Mann–Whitney test performed at each grid point.

resolution measures the intrinsic ability of the forecast system to detect (or resolve) situations in which the observed frequency of the event is different to the climatological frequency, while reliability is a measure of the bias in predicted probabilities for the event, relative to the verified event frequency. Resolution is considered the more fundamental of the two attributes, because reliability may generally be improved by a posteriori calibration of the forecast probabilities, while resolution cannot. As noted by Toth et al. (2003), the ROC score often measures only the resolution characteristics of the forecast system, whereas both reliability and resolution are measured by the BSS.

6.1. *Comparison of ROC scores*

We have calculated ROC scores at each model grid point, for presentation as global maps, and also as aggregates over the tropics, northern extratropics and Europe. Reference values for the ROC score are 0.5, the 'no-skill' value attained by random forecasts or constant forecasts of the climatological probability of the event (climatology 'forecasts') and 1.0, the score achieved by perfect deterministic forecasts (i.e. all ensemble members correctly predict the event in all years).

6.1.1. Global distribution of ROC scores. Global maps of ROC score are provided for GloSea and HadAM3 M3(ASO) predictions of 2-m temperature in the upper tercile in Figs. 3a and b. As noted in Section 4, results for M3(ASO) assess GloSea benefits in periods typically associated with the onset and subsequent development of ENSO SST anomalies in the tropical Pacific. Improved skill with GloSea in the eastern tropical Pacific is indeed evident from the greater geographical area, relative to HadAM3, with ROC scores in excess of 0.8. In tropical regions outside the Pacific Ocean, differences in ROC score are generally small, both over land and ocean regions. In the tropical Atlantic, GloSea achieves higher scores in the equatorial region off South America, while HadAM3 obtains higher scores in the southern sector south of West Africa. Over the Indian Ocean, the geographical area with ROC scores exceeding 0.7 is similar in both systems, although peak scores exceeding 0.8 are more frequent with GloSea.

Consistent with numerous other predictability studies (e.g. Graham et al., 2000), ROC scores in extratropical regions are notably lower than for the tropics with both models, particularly over continental regions. As noted from Fig. 2, the pdfs from each system differ only rarely over northern extratropical land regions, with consequently few opportunities for improved GloSea skill. Indeed, the locations of regions with relatively elevated ROC scores (e.g. >0.6) in Figs. 3a and b show a high degree of correspondence between the two systems; for example, over north-eastern North America, the North Atlantic and much of Europe. In summary, comparison of ROC scores suggests that

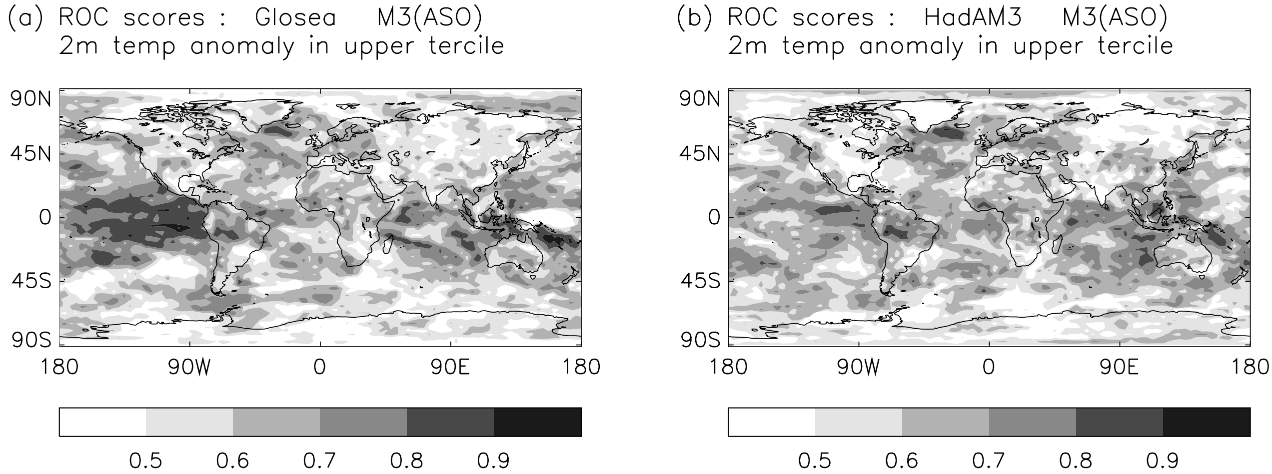

Fig 3. Global distribution of ROC scores (area under the ROC curve) for M3(ASO) 2-m temperature in the upper tercile for (a) GloSea hindcasts and (b) HadAM3 hindcasts over the 43-yr period 1959–2001. Only regions with ROC score >0.5 (the 'no-skill' threshold) are shaded. Tercile categories derived from ERA40 analyses are used for verification.

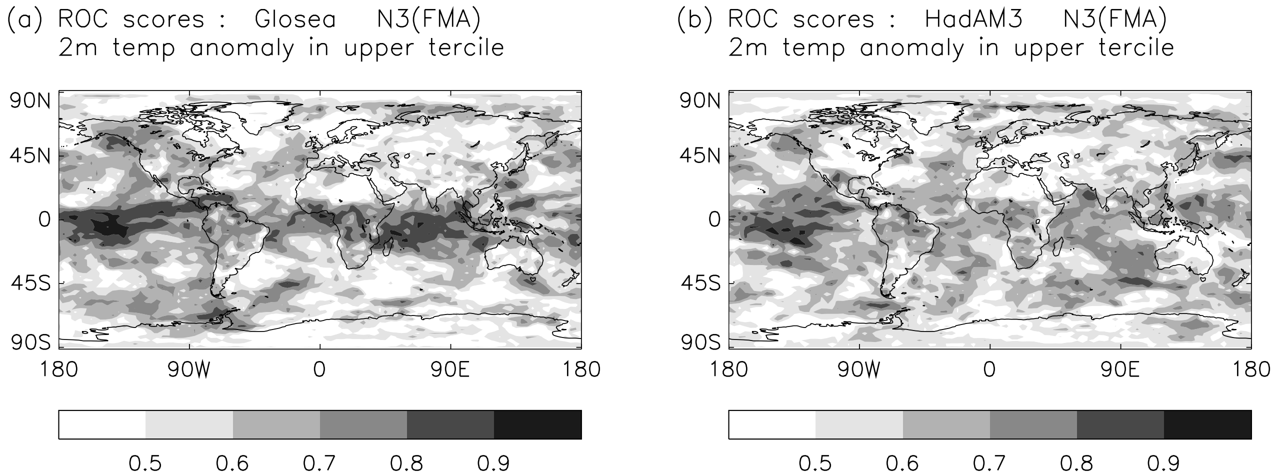

Fig 4. Same as Fig. 3, but for N3(FMA).

for the May hindcasts, HadAM3 performance is competitive with GloSea, even at three-month lead, in most areas other than in the eastern tropical Pacific, where GloSea scores appear lifted through the potential of the coupled system to predict the onset of ENSO SST anomalies.

The similarity of ROC scores in regions beyond the tropical Pacific suggests that in these regions, for M3(ASO), the two systems have similar powers of resolution. This in turn suggests that years with signals in the forecast pdfs (e.g. when the mode of the ensemble anomaly distribution is displaced from zero) tend to be similarly distributed in the time series for each system. This tendency supports the inference made in Section 5 that the frequent differences in the pdfs (Fig. 2) are more associated with a broadening of the GloSea ensemble pdfs rather than with differences in the location of the mode. As discussed later, larger ensemble spread in the GloSea system appears associated with improved reliability characteristics.

In contrast to the M3(ASO) period, comparisons for the N3(FMA) period provide assessment of GloSea benefits in periods when ENSO SST anomalies are typically declining and when lagged responses often appear in the tropical Atlantic and Indian Oceans (D02). Global maps of ROC score for GloSea and HadAM3 predictions of upper tercile 2-m temperature are provided for the N3(FMA) period in Figs. 4a and b. In the equatorial east Pacific, where ENSO SST variability is typically largest, highest scores are achieved with GloSea, the area with ROC scores exceeding 0.8 being more widespread than with HadAM3. In contrast to results for M3(ASO), differences in the GloSea and HadAM3 ROC scores in the tropical Atlantic and Indian Oceans are striking, with GloSea scores substantially higher than those for HadAM3 in both basins. GloSea scores for N3(FMA) are also higher, in these regions, than obtained in M3(ASO) (Fig. 3a), while HadAM3 scores for N3(FMA) are similar or lower than for M3(ASO) (Figs. 4b and 3b). The superior GloSea skill in the

Table 2. ROC scores (area under the ROC curve) for GloSea (G) and HadAM3 (H) probabilistic predictions of three-month-mean 2-m temperature in the upper tercile, at one-month and three-month leads, for the tropics, northern extratropics and Europe

	ROC scores											
	Tropics				Northern extratropics				Europe			
	One-month		Three-month		One-month		Three-month		One-month		Three-month	
Hindcast/period	G	H	G	H	G	H	G	H	G	H	G	H
Feb: F1(MAM), F3(MJJ)	0.74	0.71	0.67	0.65	0.63	0.63	0.57	0.59	0.65	0.64	0.60	0.62
May: M1(JJA), M3(ASO)	0.73	0.70	0.68	0.65	0.60	0.61	0.59	0.58	0.61	0.60	0.59	0.61
Aug: A1(SON), A3(NDJ)	0.74	0.71	0.70	0.65	0.60	0.61	0.58	0.57	0.58	0.59	0.60	0.59
Nov: N1(DJF), N3(FMA)	0.72	0.70	0.69	0.65	0.57	0.58	0.57	0.56	0.55	0.53	0.52	0.50

tropical Atlantic and Indian Oceans in the November hindcasts is also evident at one-month lead (not shown) and implies a strong contribution to skill from representation of lagged atmosphere–ocean responses to tropical Pacific ENSO SST anomalies. Such responses will derive from 'priming' of the 3-D ocean state with developing ENSO conditions during the assimilation, from improved ability to predict the peak amplitude of tropical Pacific SST anomalies in the first three months of the hindcast, and from associated teleconnection responses throughout the six-month period. D02 found correlations between Niño3 SST anomalies and SST anomalies in the central Indian Ocean of +0.5 at zero and six-month lags (Niño3 leading). Correlations between tropical North Atlantic and Niño3 SST anomalies were found to be lower at +0.2 and +0.4 for zero and six-month lag, respectively. Thus, the time-scale of the GloSea response, as seen through the elevated ROC scores, appears consistent with the observed time-scale (i.e. within the six-month range of the hindcast). D02 also found that a number of current CGCMs are able to reproduce the observed correlations, including HadCM3, the model on which GloSea is based. The results presented here extend those of D02 by quantifying, for a single coupled model, the benefits to atmospheric seasonal prediction skill that derive from representation of these coupled ocean–atmosphere teleconnection processes. A case study example of such benefits during the 1997 El Niño event is provided in Section 7.

Superior GloSea ROC scores are also apparent over North America in N3(FMA) (Figs. 4a and b), notably in the north-west in a region including the Gulf of Alaska and also over most central and eastern regions. The north-west region has teleconnection relationships with ENSO events and improved skill with GloSea is therefore likely to be related to improved GloSea prediction of ENSO SST anomalies in the equatorial Pacific. Elsewhere in the extratropics, the relative performance of the two systems is generally similar, although minor regional differences may be seen in Figs. 4a and b. Scores with both systems are below 0.5 over much of Europe and central Asia, notably lower than for the M3(ASO) hindcasts (Figs. 3a and b).

6.1.2. ROC scores aggregated over regions. ROC scores aggregated over the tropics, northern extratropics and Europe are provided for all hindcast sets at one-month and three-month leads in Table 2. Scores are generally above the 0.5 'no-skill' threshold in all regions even at three-month lead (although barely so for the November hindcast sets over Europe). For both systems, highest scores are found in the tropics. Here, GloSea scores are higher than HadAM3 scores in all hindcasts at both lead times, with average GloSea/HadAM3 scores over all hindcasts of 0.73/0.71 at one-month lead, and 0.69/0.65 at three-month lead. The largest percentage improvements in ROC score with GloSea (of the order of 7%) are found for the A3(NDJ) and N3(FMA) hindcasts (improvements of ~4% are found for the other hindcasts). Thus, the largest benefits over HadAM3 are found in longer-range predictions for periods associated with the growth and decay of ENSO SST anomalies and with the development of tropics-wide remote lagged responses. Regional GloSea benefits for these periods in the tropical Pacific, Atlantic and Indian Ocean regions have been discussed [using results for N3(FMA)] in the previous subsection. Note that GloSea tropics-wide scores are similar for the N3(FMA) and M3(ASO) periods (0.69 and 0.68, respectively; Table 2). However, as is evident from Figs. 4a and 3a, this similarity results from marked regional skill differences of opposite sign: scores are higher over much of the tropical Pacific in M3(ASO) relative to N3(FMA) while, as discussed in the previous subsection, tropical Atlantic and Indian Oceans scores are higher in N3(FMA). Generally, the superior GloSea results for large-scale tropical skill show that while PSSTA methods can provide competitive predictions of (local) Niño3.4 SST anomalies (e.g. the August hindcasts; Table 1), representation of coupled ocean–atmosphere processes enables better prediction of the large-scale tropics-wide conditions.

In the northern extratropics, ROC scores with both systems are similar in all hindcasts, as previously noted for the M3(ASO) and N3(FMA) periods. At one-month lead, both GloSea and HadAM3 appear to have highest skill in F1(MAM), for which scores are of the order of 0.63, and lowest skill in N1(DJF) for which scores are of the order of 0.58. At three-month lead, scores for all hindcasts drop to the order of 0.58 with both systems. ROC scores for Europe are generally similar to the northern extratropical 'averages', although scores for Europe are consistently higher (with both models at both lead times) in the February hindcasts and lower in the November hindcasts, for which scores for Europe drop to 0.55 or below (consistent with the grid-point scores of Figs. 4a and b).

6.2. *Reliability, resolution and Brier skill scores*

In this section we discuss the reliability and resolution characteristics of the GloSea and HadAM3 systems, and related BSS, calculated for the three geographical regions considered. As for the ROC score analysis we focus, for the tropical region, on the May and November hindcast sets.

6.2.1. Reliability diagrams. Forecast reliability and resolution may both be represented graphically using the reliability diagram (Wilks, 1995). Reliability diagrams are constructed by plotting the observed frequency of the event (in our case temperature in the upper tercile) calculated over subsamples stratified by forecast probability. For perfect reliability, such observed frequencies should equal the associated forecast probabilities, resulting in plotted points that lie on a diagonal.

Reliability diagrams for GloSea and HadAM3 predictions of upper tercile 2-m temperature in the tropics are provided for the May hindcasts in Fig. 5, where results for one-month lead [M1(JJA)] and three-month lead [M3(ASO)] are shown. GloSea provides substantially improved reliability over most of the probability range (GloSea points are closer to the diagonal) at both lead times. Thus, while forecast probabilities from both systems are positively correlated with the verified chance of observing the event, GloSea probabilities are better correlated. Both systems exhibit the 'overconfident' signature characterized by slopes shallower than the diagonal (i.e. when the chance of observing the event is relatively high, forecast probabilities are too large, and when the chance is relatively low, forecast probabilities are too small). The reduction in reliability with lead time (as assessed from the shallower slopes) is similar in both systems. It is striking, though, that the superiority of the GloSea performance is such that reliability at three-month lead is better than that available with HadAM3 at one-month lead, indicating substantial gains in predictable lead time.

Reliability diagrams for the November hindcasts at one-month lead [N1(DJF)] and three-month lead [N3(FMA)] are provided in Fig. 6, and show substantial GloSea performance benefits similar to those found for the May hindcasts (Fig. 5). As for the May hindcasts, superior performance with GloSea is sufficient to deliver forecasts at three-month lead that are more reliable than HadAM3 forecasts at one-month lead. GloSea reliability for the November hindcasts (Fig. 6) appears better, and more sustained with lead time, than for the May hindcasts (Fig. 5), while reliability for HadAM3 appears similar in the November sets relative to the May sets. Consequently, the performance differential is

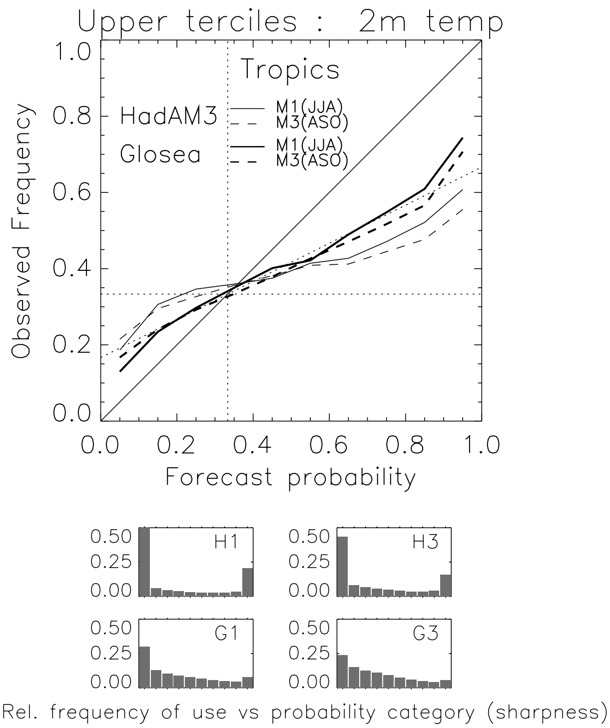

Fig 5. Reliability diagrams for tropical region (30°N–30°S) predictions of 2-m temperature in the upper tercile from May-start hindcasts over the 43-yr period 1959–2001. Tercile categories derived from ERA40 analyses are used for verification. Results are shown for GloSea (thick lines) and HadAM3 (thin lines) at one-month (solid line) and three-month (dashed line) leads. 10 probability bins are used with points plotted at bin centres, starting at 0.05 and incrementing in intervals of 0.1. For clarity of comparison, plotting of points is suppressed. The horizontal dotted line drawn at the sample frequency of the event ($\sim$0.33) is the 'line of no resolution'. The dotted line bisecting the angle between the line of no resolution and the diagonal is the 'no-skill' line. The significance of the lines of no resolution and no skill is discussed in the text. The relative frequency of use of each probability bin (sharpness) is also shown; bins are ordered left (0.05) to right (0.95). H1, H3 = HadAM3 at one- and three-month leads; G1, G3 = GloSea at one- and three-month leads.

greater, with the November hindcasts showing the largest difference in slopes. The larger GloSea reliability benefits in the November hindcasts relative to the May hindcasts are consistent with the representation in GloSea (but not in HadAM3) of lagged, remote coupled responses to ENSO events that are active over the valid period of the November hindcast set. Corresponding benefits were also found using ROC scores (Figs. 4a and b; Section 6.1).

For extratropical regions we focus on the February and November hindcasts, the sets with respectively highest and lowest ROC scores for this region at one-month lead (Table 2). Reliability diagrams for February hindcasts are provided for the northern extratropics in Fig. 7. Reliability with both systems

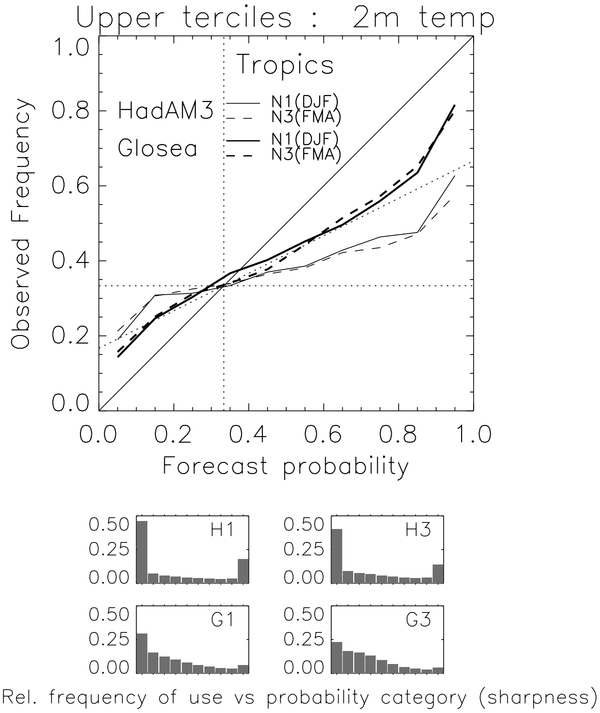

Fig 6. As Fig. 5, but for the November-start hindcasts.

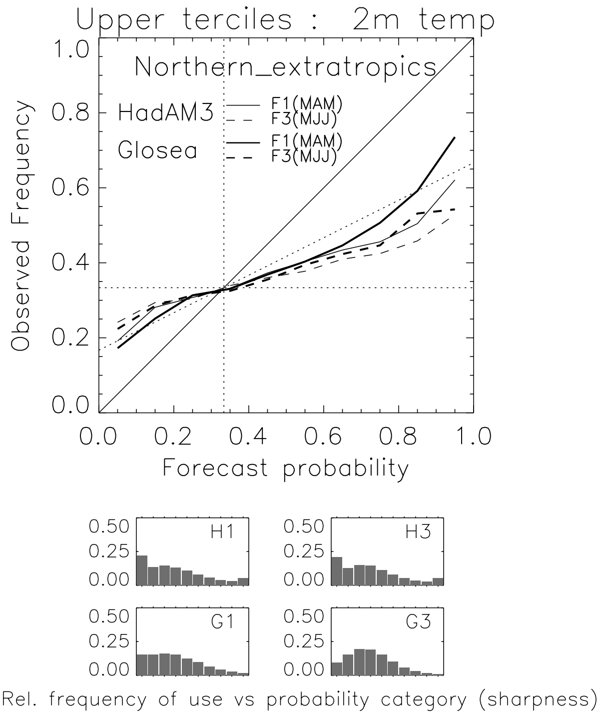

Fig 7. As Fig. 5, but for the February-start hindcasts for the northern extratropics (30°N–90°N).

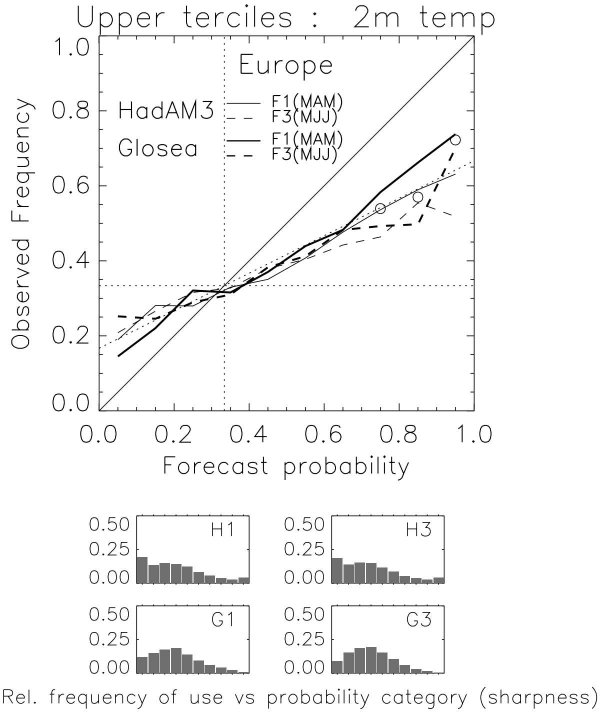

Fig 8. As Fig. 7, but for Europe (12.5°W–42.5°E, 35°N–75°N). Circles plotted for the 0.75, 0.85 and 0.95 bins show results for GloSea hindcasts at one-month lead when the years 1989, 1990 and 2000 are omitted from the sample.

is lower than for tropical regions (for which reliability of the February hindcasts is similar to that for the May hindcasts shown in Fig. 5). However, reliability is still present in the sense that increased forecast probability of the event is positively correlated with increased frequency of occurrence. As for the tropics, GloSea benefits to reliability are clearly evident at both lead times. Improvements appear most notable for forecasts indicating relatively high or low probability of the event, while improvements are marginal for forecasts in the middle probability range.

Reliability for February hindcasts for the European region (Fig. 8) is somewhat better with both systems than for the northern extratropics as a whole (Fig. 7). Slightly elevated performance for Europe relative to the northern extratropics is also evident in ROC scores for the February hindcasts (Table 2). Figure 8 shows improved GloSea reliability for European forecasts relative to HadAM3, most notably at one-month lead. As for the northern extratropics, improvements at one-month lead are particularly evident at the higher probability range (specifically, bins centred at 75%, 85% and 95% probability), and also at the lower probability range (bins centred at 5% and 15%). The sharpness diagrams indicate small samples for the higher probability bins, and therefore these points should normally be treated with caution. However, as discussed using case study results in Section 7, this indication of relatively reliable high-probability predictions from GloSea appears to have a physical basis, and therefore may not be simply an artefact of the small samples.

Corresponding diagrams for the November hindcasts (not shown) indicate reduced reliability in the northern extratropics

Table 3. Percentage GloSea benefits relative to HadAM3 for BS reliability (Rel) and resolution (Res). Results are for predictions of three-month-mean 2-m temperature in the upper tercile, at one-month and three-month leads, for the tropics, northern extratropics and Europe. Positive GloSea benefits are shown in bold type

	Reliability and resolution: per cent GloSea benefit											
	Tropics				Northern extratropics				Europe			
	One-month		Three-month		One-month		Three-month		One-month		Three-month	
Hindcast/period	Rel	Res	Rel	Res	Rel	Res	Rel	Res	Rel	Res	Rel	Res
Feb: F1(MAM), F3(MJJ)	**73.7**	**3.1**	**76.4**	**2.9**	**40.0**	**0.5**	**45.7**	−0.5	**52.6**	**0.5**	**46.2**	−0.9
May: M1(JJA), M3(ASO)	**67.4**	**3.6**	**64.7**	**2.9**	**36.1**	−1.4	**45.2**	0.0	**52.6**	**0.5**	**47.6**	−0.9
Aug: A1(SON), A3(NDJ)	**72.2**	**3.7**	**78.7**	**5.8**	**38.5**	−0.5	**55.3**	**0.5**	**61.8**	0.0	**55.0**	**0.5**
Nov: N1(DJF), N3(FMA)	**75.0**	**3.6**	**78.7**	**4.4**	**57.9**	−0.5	**50.0**	0.0	**45.5**	**0.9**	**41.3**	**0.5**

and Europe relative to the February hindcasts, with GloSea and HadAM3 reliability similar over much of the probability range. Reliability is particularly reduced for the European hindcasts. In this region, for both models at both lead times, observed frequencies for the forecast probability subsets are very similar to the climatological frequency of the event, and thus points fall near the horizontal line drawn at this frequency (referred to as the line of no resolution). Such forecasts are unable to resolve subsets of cases for which the observed event frequency is different from the overall climatological frequency. In contrast, the hindcast results presented in Figs. 5–8 all indicate degrees of resolution, for both GloSea and HadAM3 (i.e. points are vertically removed from the line of no resolution). In the next subsection we present quantitative comparisons of resolution and reliability for GloSea and HadAM3, calculated over all forecast probability bins.

6.2.2. Brier score reliability and resolution. Numerical measures of reliability and resolution, aggregated over all forecast probability bins, form two 'decomposition' terms in the BS (Murphy, 1973; see also Wilks, 1995). The BS reliability term ($B_{\rm rel}$) consists of a weighted average of the squared vertical distances between the diagonal and the plotted points, with smaller values indicating better reliability. The BS resolution term ($B_{\rm res}$) is given by a weighted average of the squared vertical distance of the plotted points from the line of no resolution, with better resolution indicated by larger values. In each case, the weighting is by the frequency of use of the probability bin (sharpness; see lower panels in Figs. 5–8). The third term in the BS decomposition, referred to as the uncertainty, is independent of the forecast system and given by $U = \bar{o}(1 - \bar{o})$, where $\bar{o}$ is the observed climatological frequency of the event. Thus

$$\mathrm{BS} = B_{\rm rel} - B_{\rm res} + U, \quad (1)$$

where $0 \leq \mathrm{BS} \leq 1$. The BS is negatively oriented such that a perfect forecast system would attain BS = 0.

GloSea benefits over HadAM3 for both $B_{\rm rel}$ and $B_{\rm res}$ may be assessed using the skill score formula

$$\mathrm{SS} = \frac{B_{\rm G} - B_{\rm H}}{B_{\rm perf} - B_{\rm H}} \times 100\%, \quad (2)$$

where $B_{\rm G}$ and $B_{\rm H}$ are the GloSea and HadAM3 values, respectively, and $B_{\rm perf}$ is the value attained by a perfect forecast system. For reliability, $B_{\rm perf} = 0$, while for resolution $B_{\rm perf} = \bar{o}(1 - \bar{o})$, the uncertainty (as is clear from substituting $\mathrm{BS} = 0$, $B_{\rm rel} = 0$ in eq. 1).

In terms of the unweighted geometrical distances of plotted points from the no-skill and no-resolution lines, the reliability diagrams (e.g. Figs. 5 and 6) suggest that for each bin GloSea improvements in reliability are matched by numerically equal improvements in resolution. However, Table 3 shows that, when weighting is included, GloSea benefits to reliability are substantially larger than for resolution. The impact of the frequency-of-use weighting may be appreciated by noting the much stronger weighting in the HadAM3 hindcasts to the outer probability bins, where in both systems reliability is worst but resolution is best. The difference in weighting therefore acts to amplify GloSea advantages for overall reliability, but diminish advantages to overall resolution.

In tropical regions, substantial GloSea benefits to reliability (65–80%) are found in all hindcast sets at both lead times. Improvements in resolution are smaller, at up to 6%. As found for the ROC scores, the best GloSea improvements in both reliability and resolution occur at three-month lead in the August and November hindcasts, suggesting that improvements are general but are enhanced by GloSea representation of developing and decaying ENSO events and related tropics-wide responses.

In the northern extratropics, GloSea benefits to reliability are smaller than for the tropics but still substantial (order 35–60%; Table 3). With the exception of the November hindcasts, improvements appear somewhat larger for Europe than for the northern extratropics as a whole, particularly at one-month lead. In contrast, GloSea benefits to resolution in the northern extratropics and Europe are generally neutral, being small in magnitude and with positive and negative signs (nearly) equally frequent. The lack of GloSea improvement to resolution outside the tropics is consistent with the ROC scores, which are similar with both systems for the northern extratropics and Europe (Table 2). The consistency follows from the previously mentioned observation of Toth et al. (2003) that, under the condition that the reliability curve is monotonically increasing (as is true in Figs. 5–7), the ROC score is independent of reliability and measures only the resolution characteristics of the forecast system (see also Wilks, 2001).

In summary, results suggest that in the tropics GloSea brings improvements both to the 'detection' of the event (ROC score/resolution) and the calibration (reliability) of the forecast event probability. In the extratropics, on the other hand, GloSea brings little or no improvement in event detection, but still improves calibration of the forecast probabilities.

The above analysis suggests that cases of significant differences in HadAM3 and GloSea forecast pdfs (Fig. 2) are mainly associated with improved reliability in the GloSea system, while differences related to improved resolution and ROC score (and thus improved event 'detection') are fewer and mainly confined to the tropics. Comparison of HadAM3 and GloSea sharpness diagrams for the tropics (Figs. 5 and 6) suggests that improved reliability with GloSea is due, in part, to broader ensemble spread. Indeed, the HadAM3 hindcasts appear unrealistically sharp, with the high frequency of use of the highest and lowest probability bins indicating a quasi-deterministic, closely-grouped behaviour of the ensemble. This quasi-deterministic behaviour is much reduced in the corresponding sharpness plots for the GloSea forecasts, which are more evenly distributed over the probability categories. Improved GloSea spread and reliability must derive from representation of initial ocean uncertainties (absent in HadAM3) and, via coupling, better representation of the evolving ocean–atmosphere uncertainty. However, the quasi-deterministic nature of the HadAM3 ensemble, particularly in the tropics, suggests that part of the GloSea improvements to reliability might be matched by employing a posteriori methods to broaden the HadAM3 spread. Such broadening might be achieved, for example, through ensemble dressing techniques based on historical error statistics (e.g. Roulston and Smith, 2003), and would likely add more benefit to the HadAM3 ensembles than to the (already dispersive) GloSea ensembles. However, such methods add no further physical information to the forecast process, and therefore improvements to 'event detection' (resolution) are likely to be limited. We note also that improvement to reliability, although not resolution, might be achieved by simple calibration or bias correction of the HadAM3 forecast probabilities. This method effectively uses probability bias information, available from the hindcast performance, to move points horizontally towards the diagonal. Thus, while GloSea benefits to reliability may be reproduced, to some extent, by post-processing of HadAM3 output, GloSea gains to resolution in the tropics represent unique benefits from ocean–atmosphere coupling.

6.2.3. Brier skill scores. The BSS, with climatology as reference, is given by

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{cl}}}, \tag{3}$$

where $\mathrm{BS}_{\mathrm{cl}}$ is the BS awarded to climatology forecasts. A positive BSS indicates the forecast system is more skilful (as measured by the BS) than climatology forecasts, and a perfect deterministic forecast system would achieve the upper bound of BSS = 1. We note that, by definition, climatology forecasts have perfect reliability but no resolution ($B_{\mathrm{rel}} = B_{\mathrm{res}} = 0$) and are represented on reliability diagrams by a single point at the intersection of the perfect reliability and no-resolution lines. Consequently, $\mathrm{BS}_{\mathrm{cl}} = U$, and following, for example, Wilks (1995), eqs. (1) and (3) give

$$\mathrm{BSS} = \frac{B_{\mathrm{res}} - B_{\mathrm{rel}}}{U}. \tag{4}$$

Equation (4) shows that use of climatology as a reference 'predisposes' the BSS towards negative values, because most systems will have $B_{\mathrm{rel}} > 0$ (i.e. reliability worse than the perfect, although 'useless', reliability of climatology forecasts). This artefact of the BSS tends to mask capability for resolution, because forecasts with $B_{\mathrm{res}} > 0$, and therefore some ability (unlike climatology forecasts) to resolve situations when the event probability is different from the climatological probability, will nevertheless be awarded a negative BSS if $B_{\mathrm{res}} < B_{\mathrm{rel}}$. For this reason, BSS results tend to give a more pessimistic assessment than ROC scores which, as mentioned, often measure only the resolution characteristics of the system.

A summary of BSS for the tropics, northern extratropics and Europe for all four hindcast sets at both lead times is provided in Table 4. Positive BSS, indicating better performance than climatology, are achieved only with the GloSea system and are almost exclusively confined to the tropics. This contrasts with the ROC score results of Table 2, which indicate improved performance relative to climatology (ROC score >0.5) in most regions and lead times. As discussed above, the harsher BSS assessments are an artefact of the reference against the perfect (but useless) reliability of climatology forecasts. In the tropics, GloSea achieves positive scores in all hindcasts and lead times. Consistent with the reliability diagrams (Figs. 5 and 6) GloSea scores at three-month lead are better than HadAM3 scores at one-month lead. The only positive BSS outside the tropics is achieved with the GloSea model for the F1(MAM) hindcasts over Europe (for which BSS = 0.03). Consistent with the reliability and resolution results of Table 3, the largest GloSea improvements in BSS are found in the tropics for the A3(NDJ) and N3(FMA) hindcasts. Comparison of Tables 2 and 4 indicates that while ROC scores with the two systems are comparable outside the tropics, BSS are better with the GloSea system in all regions. This result is consistent with the finding (Table 3) that GloSea improvements outside the tropics are essentially found only in the reliability

Table 4. BSS for GloSea (G) and HadAM3 (H) predictions of three-month-mean 2-m temperature in the upper tercile, at one-month and three-month leads, for the tropics, northern extratropics and Europe. Positive scores are shown in bold type

	Brier skill scores											
	Tropics				Northern extratropics				Europe			
	One-month		Three-month		One-month		Three-month		One-month		Three-month	
Hindcast/period	G	H	G	H	G	H	G	H	G	H	G	H
Feb: F1(MAM), F3(MJJ)	**0.12**	−0.03	**0.03**	−0.18	−0.01	−0.06	−0.07	−0.13	**0.03**	−0.02	−0.04	−0.08
May: M1(JJA), M3(ASO)	**0.10**	−0.07	**0.02**	−0.16	−0.07	−0.12	−0.05	−0.12	−0.05	−0.14	−0.02	−0.06
Aug: A1(SON), A3(NDJ)	**0.13**	−0.02	**0.08**	−0.14	−0.05	−0.14	−0.06	−0.16	−0.04	−0.13	−0.01	−0.07
Nov: N1(DJF), N3(FMA)	**0.11**	−0.06	**0.07**	−0.14	−0.09	−0.15	−0.07	−0.15	−0.09	−0.19	−0.12	−0.21

component of performance, which is measured by BSS but not (under conditions noted by Toth et al., 2003) by the ROC score.

The BSS and ROC scores generally agree on the relative skill levels found for the four hindcast sets. Agreement in scores is best for the extratropics and Europe at one-month lead. Here, highest scores with both measures are found, for both systems, in the February hindcasts, while lowest scores are consistently found in the November hindcasts. The latter result is consistent with the relatively poor reliability and resolution characteristics noted for the November hindcasts from reliability diagrams (Section 6.2.1).

From eq. (4) a ‘no-skill’ line (BSS = 0), bisecting the angle between the perfect reliability and no-resolution lines, may be drawn on the reliability diagrams (as in Figs. 5–8). Forecast subsamples with points located in the area bounded by the no-skill line and the vertical line drawn through the climatological probability of the event contribute positively to the BSS. Points located outside this ‘skilful area’ contribute negatively. The location of the plotted points relative to the skilful area allows some interpretation of the BSS of Table 4. In the tropics (Figs. 5 and 6), for example, plotted points for GloSea frequently lie near or within the skilful area, the latter contributing to a positive BSS. In contrast, HadAM3 points invariably lie outside the skilful area (indicating negative BSS). For the northern extratropics (Fig. 7) points with both systems lie generally outside the skilful area at both lead times. Results for F1(MAM) over Europe (Fig. 8) show that the positive BSS achieved by GloSea (BSS = 0.03; Table 4) results from positive contributions from the previously mentioned forecasts in the 75%, 85% and 95% bins (and also from weaker but more numerous contributions from forecasts in the 5% and 15% probability bins).

7. Case study investigations

In this section we present case study investigations to add illustrative detail to results with long-term skill scores presented in Section 6.

7.1. Comparison of tropics-wide responses to the El Niño event of 1997

The improved skill with GloSea in N3(FMA) over the tropical Atlantic and Indian Oceans, discussed in Section 6 (Figs. 4a and b), is also a marked feature of corresponding ROC maps for the August hindcasts (not shown), both at one-month and three-month leads. To investigate further the association of this improved skill with responses to ENSO activity, an annual breakdown of the percentage area for which GloSea and HadAM3 pdfs of 2-m temperature anomaly are significantly different has been performed for each of the three ocean basins (using the Wilcoxon–Mann–Whitney test). The resulting time series for A3(NDJ) is shown in Fig. 9. As a measure of coincident and preceding ENSO activity the OND periods with largest magnitude Niño3.4 indices (nine negative, nine positive) are annotated in Fig. 9. For the Indian Ocean (right-hand bars) there is a clear association of large percentage area with significant differences in the pdfs and higher ranking ENSO events, indicating a coupled response in GloSea that is not supported in HadAM3. The association is particularly evident for the five strongest El Niño events. The main exception is the peak in NDJ 1983/1984, for which there is no associated high-ranking ENSO event in OND (although strong El Niño conditions were present in April–June of 1983). There is some evidence of a similar, although much weaker, association in the tropical Atlantic (e.g. peaks in 1972/1973, 1982/1983 and 1987/1988) consistent with a smaller response to ENSO SST anomalies found for that region by D02.

A case study of GloSea and HadAM3 performance for the August 1997 hindcasts is provided in Fig. 10, and gives an example of the practical benefits of ocean–atmosphere coupling, relative to two-tier systems, for seasonal prediction of 2-m temperature throughout the tropical belt. Over oceans, HadAM3 ensemble mean 2-m temperature anomalies for A1(SON) (Fig. 10a) will reflect observed SST anomalies for July. Use of PSSTA results in a reasonable HadAM3 representation of the observed El Niño

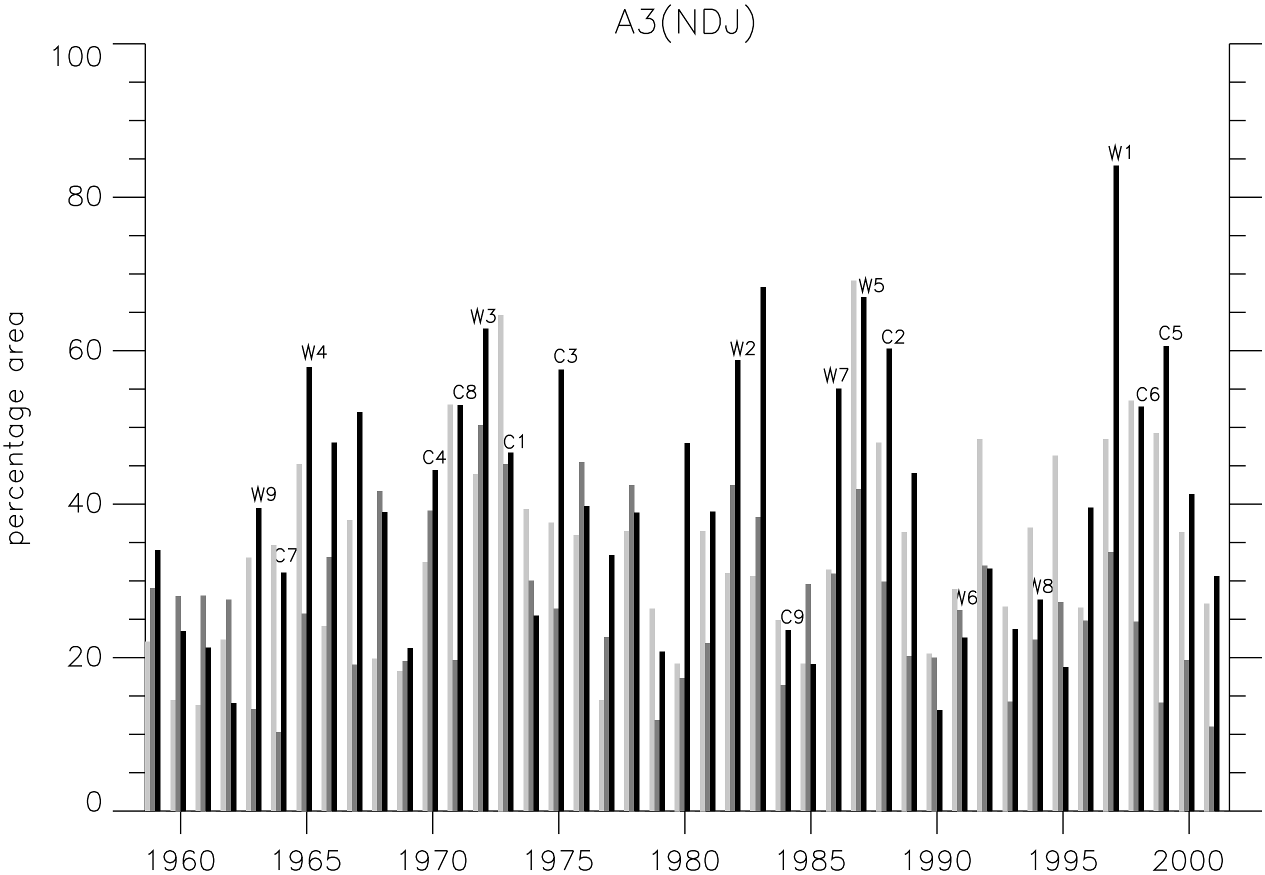

Fig 9. Time series of the percentage area over tropical oceans for which GloSea and HadAM3 ensemble distributions for A3(NDJ) 2-m temperature anomalies are found significantly different at the 95% level (from the Wilcoxon–Mann–Whitney test). Light (left-hand) bars denote tropical Pacific; darker (middle) bars denote tropical Atlantic; dark (right-hand) bars denote Indian Ocean. The labelling of years refers to November and December of the NDJ period. The corresponding OND periods with the largest magnitude Niño3.4 indices (nine negative, nine positive) are annotated. The indices, taken from the International Research Institute (IRI) internet site (http://iri.columbia.edu/climate/ENSO/background/pastevent.html) are defined using a five-month running average of the Niño3.4 index (Trenberth, 1997). W = El Niño, C = La Niña, and following numbers give the ranked order (separately for negative and positive signs) of the Niño3.4 index over the 1959–2001 period.

pattern of warm anomalies in the tropical east Pacific at one-month lead (SON; Figs. 10a and e), although the strategy of damping persistence with lead time results in the erroneous decline of the warm anomalies by NDJ (Figs. 10b and f). In contrast, GloSea maintains the tropical Pacific anomalies into the NDJ period (Fig. 10c and d), although warm anomalies extend too far west at both lead times. In the tropical south Atlantic, HadAM3 predicts a cold SON 2-m temperature anomaly (Fig. 10a) as a result of an observed cold SST anomaly in July, and this is persisted (with some damping) into the NDJ period (Fig. 10b). In contrast, GloSea predicts the disappearance of the cold anomaly by SON and further anomalous warming by NDJ (Figs. 10c and d), predictions that verify well with the observed evolution (Figs. 10e and f). The Indian Ocean also undergoes a broad-scale warming over the SON and NDJ periods (Figs. 10e and f). Here the general pattern of warm anomalies is again well modelled by GloSea (Figs. 10c and d), although the predicted warming is somewhat too widespread. In contrast, use of damped PSSTA results in erroneous HadAM3 predictions of a cooling rather than warming trend in 2-m temperature over the Indian Ocean (Figs. 10a and b). The improved GloSea predictions over the tropical Atlantic and Indian Oceans described above derive from coupled ocean–atmosphere responses that are not supported by the HadAM3 system. Moreover, elevated GloSea ROC scores in these regions (similar to those noted for the November hindcasts in Fig. 4) indicate that such improved GloSea prediction relative to HadAM3 is typical in years with ENSO events.

7.2. *Case study for Europe: NAO related spring season predictability*

The F1(MAM) period was chosen for case studies over Europe because of the relatively high skill (with both systems) in this season (Tables 2 and 4). A time series of the percentage European

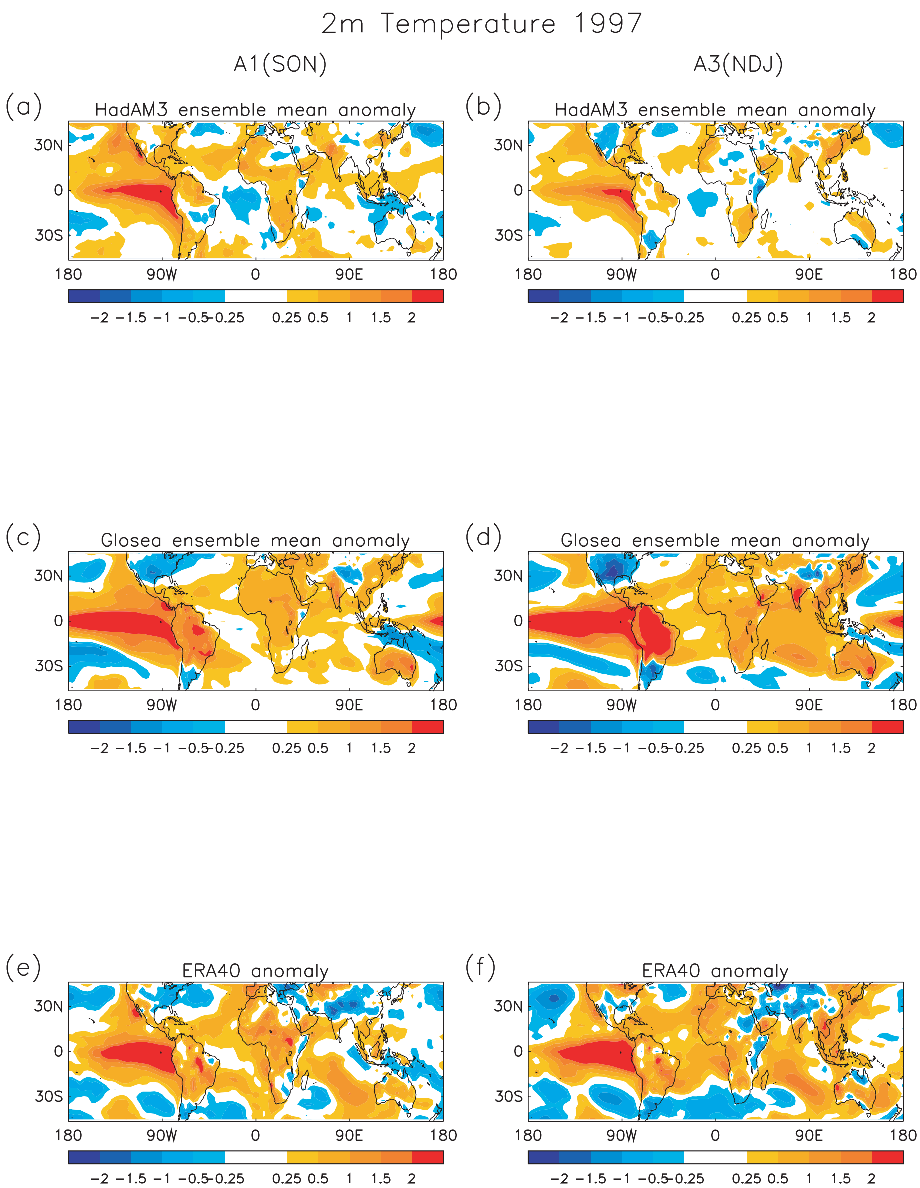

Fig 10. HadAM3 and GloSea ensemble mean 2-m temperature anomalies for A1(SON) 1997 and A3(NDJ) 1997/1998 and corresponding verification: (a) HadAM3 ensemble mean A1(SON) 1997; (b) HadAM3 ensemble mean A3(NDJ) 1997/1998; (c) as (a) but for GloSea; (d) as (b) but for GloSea; (e) observed (ERA40) anomaly for SON 1997; (f) observed (ERA40) anomaly for NDJ 1997/1998. Blue shading indicates anomalies <-0.25 K and red shading indicates anomalies >0.25 K.

area, for which the GloSea and HadAM3 pdfs for F1(MAM) 2-m temperature anomaly are significantly different (similar to that presented for tropical oceans in Fig. 9), revealed that years with large areas of pdf differences, representing 'opportunities' for improved GloSea skill, are often associated with years of high NAO index, either in MAM or in the preceding DJF season. This finding raised the question of whether the superior GloSea reliability and BSS for the F1(MAM) hindcasts of upper tercile 2-m temperature over Europe, discussed in Section 6.2, result from improved GloSea predictions in years with relatively large positive NAO index. A time series breakdown of GloSea and HadAM3 BSS is provided in Fig. 11. We shall refer to the BSS for an individual year as BSS_y, to distinguish it from BSS (the score calculated over the whole 43-yr sample). Figure 11 shows three years, 1989, 1990 and 2000, in which GloSea BSS_y are exceptionally high (order 0.6) and improved relative to HadAM3. GloSea scores are not invariably better than HadAM3 (e.g. 1984 and 2001), but because we wish to identify a physical basis for the improved overall GloSea performance, we focus here on the three years with high GloSea BSS_y.

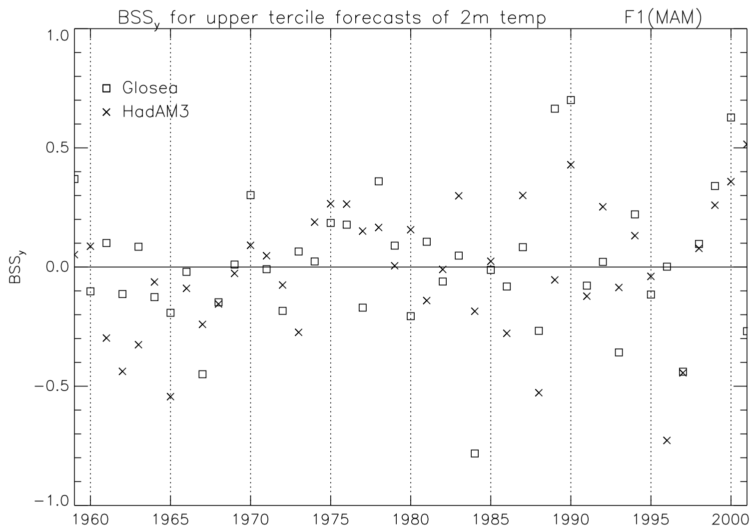

Fig 11. Time series of Brier skill scores (BSS_y) for GloSea (squares) and HadAM3 (crosses) probability forecasts of F1(MAM) 2-m temperature over Europe in the upper tercile. BSS_y are calculated separately for each year aggregating over all grid points in the European region, with constant prediction of the climatological frequency of the event (1/3) used as the reference forecast strategy. Tercile categories derived from ERA40 analyses are used for verification.

Table 5. Averaged three-month NAO index (NAOI) (from data available on the University of East Anglia internet site, http://www.cru.uea.ac.uk/cru/data/nao.htm; see also Jones et al., 1997) for three selected years discussed in the text. The ranked order in the 1959–2000 series is also shown (positive and negative indices are ranked separately)

	DJF		MAM	
Year	NAOI	Rank	NAOI	Rank
1989	3.00	1	1.04	7
1990	2.13	4	1.23	5
2000	2.28	3	−0.83	10

The 1989, 1990 and 2000 MAM seasons were each associated with a high-ranking positive NAO index in the preceding DJF period (ranks 1, 4 and 3 in the 1959–2000 series; Table 5). In 1989 and 1990, the high positive index persisted into the MAM period itself. In 2000, a strong negative index in April resulted in a negative MAM average, although the March and May indices were individually both positive (at 0.54 and 0.31, respectively). Consistent with positive NAO signatures in at least two of the three months, each of the three MAM seasons was associated with warm observed 2-m temperature anomalies over much of Europe. Figure 11 indicates that GloSea predictions in these three years contribute strongly to the overall positive GloSea BSS for Europe (Table 4) and, by inference, to the superior GloSea reliability at one-month lead indicated in Fig. 8 and Table 3 (BS_{rel} is improved by 52.6%). This was confirmed by recalculating the BSS and reliability diagrams omitting the three years in question. The resulting BSS is reduced from +0.03 to −0.01, the latter value being similar to that achieved with HadAM3 over the full sample (Table 4). Consistent with the reduced BSS, there is a marked reduction of reliability for the 75% and 85% probability bins, such that these points are approximately collocated with HadAM3 points (see additional plotted points on Fig. 8), although the reliability of the 95% probability forecasts is little changed.

Thus, improved skill with GloSea relative to HadAM3 for upper tercile 2-m temperature in the F1(MAM) period appears partly attributable to very successful GloSea forecasts in a few years associated with a high-ranking positive NAO index in the preceding winter (DJF) period, that remains generally positive into spring. Figure 11 indicates that GloSea provides the best improvement over HadAM3 in MAM 1989, and verification of the deterministic (ensemble mean) forecasts for this case is provided, as an example, in Fig. 12. The ERA40 analysis (Fig. 12c) shows a typical positive NAO signature of warm anomalies over Scandinavia and eastern central Europe (of order 1.5K). The anomalies are well matched in the GloSea ensemble mean (Fig. 12b), both in magnitude and location, while in HadAM3 (Fig. 12a) the positive anomaly is too weak and centred west of the observed position, and an erroneous negative anomaly centred over the Black Sea is more pronounced than with GloSea.

Previous studies on the association of North Atlantic SST anomalies with NAO variability (e.g. RF03) suggest that the improved performance with GloSea described above may be attributable to improved SST anomaly prediction resulting from representation of coupled ocean–atmosphere processes. RF03 have demonstrated the role of a North Atlantic 'tripole' SST anomaly pattern in reinforcing winter and spring season NAO activity. The tripole pattern associated with a positive NAO index (referred to here as the positive tripole pattern) comprises negative SST anomalies in the subtropical North Atlantic and to the south of Greenland, and positive anomalies off the east coast of North America. The same pattern, with reversed polarity, is associated with negative NAO index. Improved GloSea 2-m temperature predictions over Europe might be expected in cases when development or intensification of the tripole SST anomaly pattern occurs subsequent to January, and is well

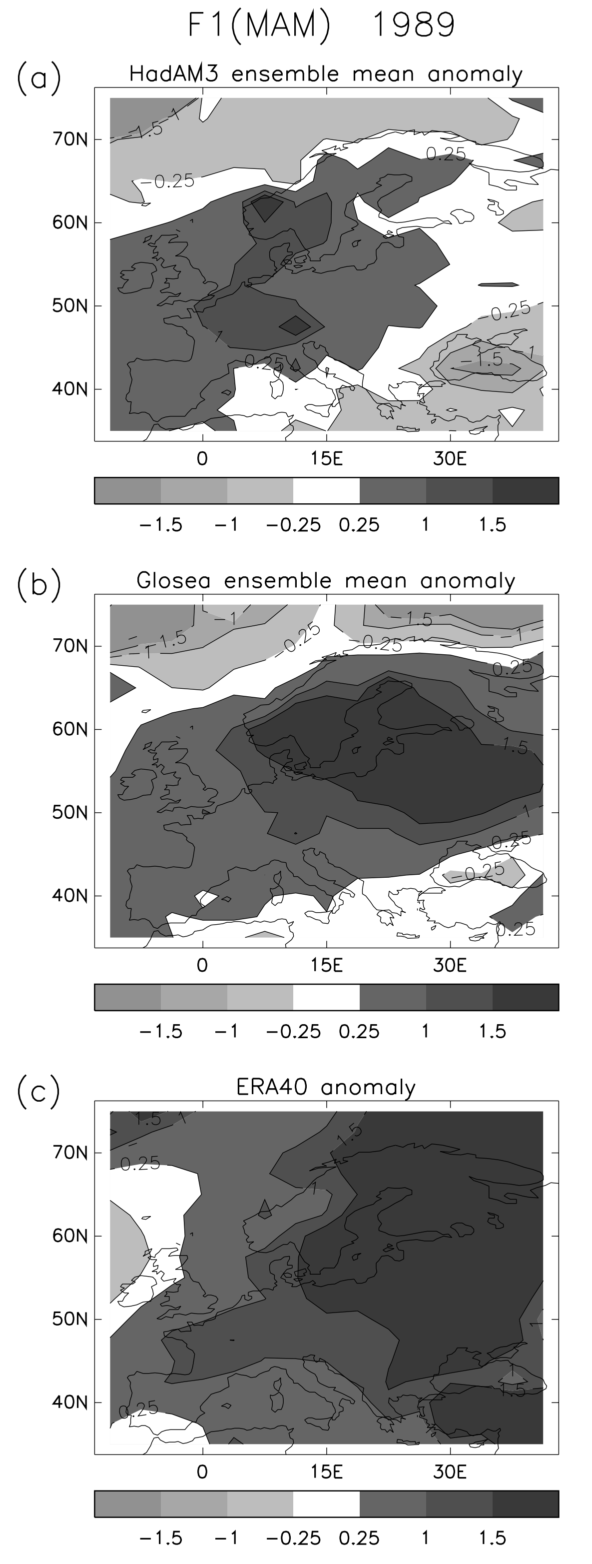

Fig 12. (a) HadAM3 ensemble mean anomaly for F1(MAM) 1989 2-m temperature over Europe. (b) As for (a) but for GloSea. (c) ERA40 2-m temperature anomalies for MAM 1989. Light shading indicates anomalies <− 0.25 K and dark shading indicates anomalies >0.25 K.

predicted by GloSea over the February–May period. In such cases, the appropriate forcing pattern will be present in GloSea, but absent in HadAM3, which is forced with persisted January SST anomalies. Subjective assessment of GloSea SST anomaly predictions for the three years considered supports the above reasoning. The most striking case is that of 1989, also the case with the most striking GloSea improvements in European 2-m temperature predictions (Fig. 11). The observed and GloSea predicted evolutions of North Atlantic SST anomalies in this case are briefly discussed below.

Observed monthly-mean SST anomalies in January 1989 (Fig. 13a) show organized structure in only one pole of the tripole pattern, in the subtropical North Atlantic region near 15°N. There is no evidence of an organized warm anomaly off the east coast of North America, as would be consistent with reinforcing of a positive NAO; in fact, patches of cold anomaly predominate at this time. However, by February (Fig. 13b) the two remaining poles in the (positive) pattern have developed (i.e. warm SST anomalies off the eastern coast of North America and cold SST anomalies south of Greenland). By May (Fig. 13c) the pattern shows further amplification in all three centres. The key features of the observed anomalies are reproduced in the GloSea ensemble-mean prediction for February (Fig. 13d), although there are some errors in positioning and geographical extent of the anomalies, and a region of cold SST anomalies extends too far southward east of Nova Scotia. By May (Fig. 13e) organization of the warm anomaly off the eastern coast of North America and amplification near 40°N appears well predicted, although the geographical extent of the anomalies is too small. There is also evidence of a separate warm anomaly centre further north-east, as observed. The cold anomaly south-east of Greenland is well represented, although the erroneous cold anomaly east of Nova Scotia is still present. Thus, GloSea SST forcing on the atmosphere develops more or less as observed, with intensification of the tripole anomaly pattern associated with reinforcing a positive NAO index. In contrast, SST forcing of HadAM3, derived from January SST anomalies (Fig. 13a), remains static and contains only the subtropical North Atlantic pole of the tripole pattern.

The 1990 and 2000 cases (not shown) are different in character to the 1989 case in the sense that the positive tripole pattern was already largely established in the observed January SST anomalies. However, in both years further development or organization of the tripole pattern takes place subsequent to January. GloSea predictions of these further developments, although imperfect, result in improved representation of the SST anomaly forcing relative to persistence of January SST anomalies. In these two years, the presence of the tripole anomaly pattern in January is consistent with HadAM3 obtaining better BSS_y than obtained for 1989, and with the smaller differential in BSS_y relative to GloSea (Fig. 11).

The case study evidence presented supports the hypothesis of RF03 that North Atlantic SST anomalies play a key role in reinforcing NAO activity. It also emphasizes the role of coupled

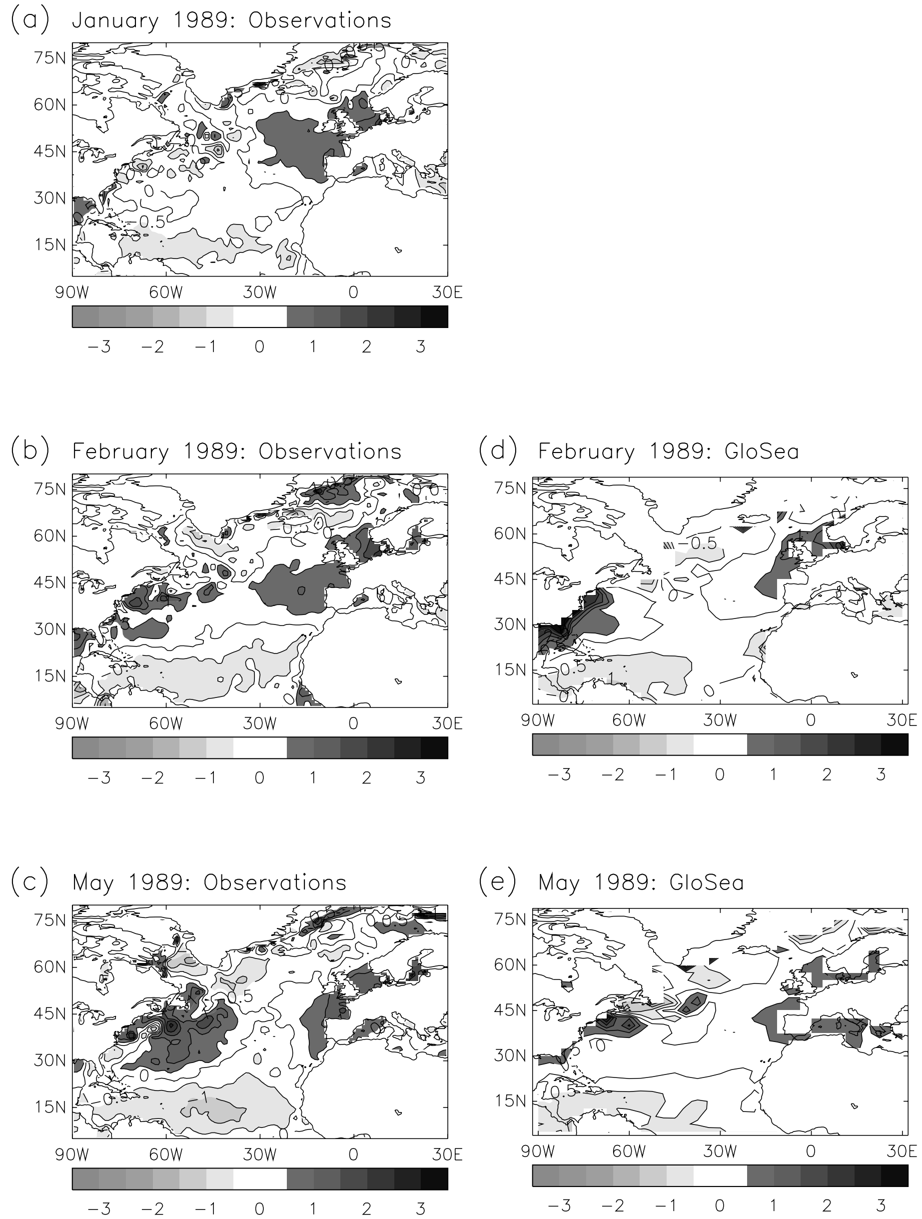

Fig 13. Observed monthly-mean North Atlantic SST anomalies (HadISST1) and corresponding ensemble mean values from the GloSea hindcast started 1 February 1989. (a) January 1989 observed anomalies [persisted in HadAM3 F1(MAM) hindcasts], (b) February 1989 observed anomalies, (c) May 1989 HadiSST1 observed anomalies, (d) February 1989 GloSea hindcast ensemble-mean anomalies, and (e) May 1989 GloSea hindcast ensemble-mean anomalies. Contour interval is 0.5 K. Light shading indicates anomalies <-0.5 K and dark shading indicates anomalies >0.5 K.

ocean–atmosphere processes in mutually sustaining the tripole SST anomaly pattern and the NAO signature. In the 1989 case, the sequence suggests an ocean response to strong positive winter NAO, which results in a tripole SST anomaly pattern in February. The improved skill of GloSea relative to HadAM3 for the MAM season indicates that the tripole pattern aids maintenance of the positive NAO into the spring period. Importantly, the case studies present evidence that coupled models can predict development of the tripole pattern and subsequent positive feedback/phase-locking with the NAO response in the atmosphere. Thus, the result is very encouraging in terms of prospects for improving seasonal predictions in the European region.

Results for prediction of F1(MAM) temperature over Europe in the lower tercile (not shown) indicate generally lower skill, for both models, relative to predictions for the upper tercile. GloSea performance is again superior, but less marked than for the upper tercile. This result is reminiscent of the findings of Davies et al. (1997) who demonstrated, for MAM, a clear association between strong positive tripole SST patterns and positive NAO index, but a much weaker reciprocal relationship between strong negative tripole patterns and negative NAO index. The same study also found only weak associations between strong DJF tripole SST patterns of either sign and simultaneous NAO index. Weak association between SST anomaly pattern and NAO phase in DJF implies that predictions of the North Atlantic SST evolution over that season will provide little advantage to model forecasts of the NAO and its impacts (e.g. on 2-m temperature). This is consistent with the notably lower performance found here for N1(DJF) relative to F1(MAM) with both GloSea and HadAM3 (Tables 2 and 4).

8. Summary and discussion

The performance of the Met Office's ocean–atmosphere CGCM seasonal prediction system (GloSea) has been compared with a former AGCM+PSSTA system (HadAM3) using 43-yr hindcast data sets generated as part of the EU project DEMETER. The GloSea and HadAM3 systems share the same atmospheric component, and thus differences in prediction skill may be attributed solely to the influence of ocean–atmosphere coupling in GloSea, which impacts both on the forecast evolution and on the ensemble perturbation strategy. The GloSea and HadAM3 configurations are typical of other coupled and forced GCM systems used in operational seasonal prediction, and the performance comparison is therefore of general interest in assessing the relative benefits of the two types of system. We expect that benefits, albeit generally smaller in magnitude, would also be found relative to many AGCM+FSST systems because, where statistical SST prediction is used, FFST systems (as PSSTA systems) cannot directly support the remote coupled responses shown here to benefit GloSea skill in the tropics. Analysis has focused on global comparisons of skill for probabilistic prediction of 2-m temperature in the upper tercile, and on selected case study comparisons for the tropics and Europe. Overall, strong evidence has been presented that coupled models can provide substantial benefits for seasonal prediction not only in tropical regions, but also in the extratropics. Results are summarized below.

All performance measures indicate the largest and most consistent benefits for GloSea relative to HadAM3 in the tropics. Here, improvements result from enhanced resolution (essentially 'event detection') and from enhanced reliability (calibration of the forecast probabilities). Percentage benefits to reliability (order 65–80%) are markedly larger than benefits to resolution (up to 6%). The most notable advantages are found at three-month lead in the August and November hindcasts [A3(NDJ) and N3(FMA), respectively]. Evidence has been presented that these marked benefits are associated with improved long-lead GloSea prediction of developing and decaying ENSO SST anomalies, and also from representation of lagged coupled ocean–atmosphere teleconnection effects to ENSO not represented in HadAM3. This latter benefit has been illustrated with a case study of improved GloSea prediction of warming in the tropical Atlantic and Indian Ocean regions associated with the 1997 El Niño.

Skill in extratropical regions is lower than for tropical regions with both systems though, as for the tropics, GloSea provides superior performance. GloSea benefits appear to derive from improved reliability (percentage improvements are of order 35–60%) rather than resolution. Measured over large geographical regions, the two systems appear to have similar powers of resolution. This is supported by the global distributions of ROC score, which are generally similar, outside the tropics, with both systems (recall the ROC score tends to measure only forecast resolution). An exception is the improved GloSea ROC scores over the North Pacific and North America in the November hindcasts at three-month lead [N3(FMA)]. These improvements in 'event detection' are likely to be related to teleconnections with ENSO SST anomalies which are better predicted with GloSea.

Although GloSea improvements to resolution appear mainly restricted to tropical regions, the universal improvement in reliability, indicating better correlation of 'issued' forecast probabilities with the frequency of event occurrence, is also a useful benefit for operational forecasting. It has been noted that part of the improvement in reliability is likely to be a reflection of the 'quasi-deterministic' behaviour of the HadAM3 ensemble, a weakness that could be addressed by post-processing. However, post-processing of the HadAM3 ensemble would not bring the improvements to resolution attained by GloSea in tropical regions.

For the northern extratropics and Europe, seasonal dependence of skill is most consistent at one-month lead where, for both systems, ROC scores and BSS are highest in F1(MAM) and lowest in N1(DJF). Improved GloSea reliability and BSS over Europe in F1(MAM) have been shown to derive largely from a few very successful forecasts in years with strong positive NAO index in the preceding DJF season. Case study evidence has been presented which suggests that improved GloSea predictions in these cases are attributed to representation of coupled ocean–atmosphere processes associated with mutual sustaining/reinforcement of the North Atlantic SST anomaly tripole pattern of RF03 and the associated NAO signature in the atmosphere. It has been speculated that the higher prediction skill found with both systems for the spring season relative to the winter season may be related to the apparent stronger association of the tripole pattern and NAO phase in spring, as evidenced in Davies et al. (1997).

9. Acknowledgments

The work of the ECMWF project team in project coordination, and in development of the DEMETER archive and website, is gratefully acknowledged. The authors are grateful for help and discussions with Met Office colleagues Malcolm MacVean, Adam Scaife, Ken Mylne and Neill Bowler. This work was supported by the European Framework 5 Programme under contract EVK2-CT-1999-00024.

References

Barnston, A. G., Mason, S.J., Goddard, L., DeWitt, D. G. and Zebiak, S. E. 2003. Multi-model ensembling in seasonal climate forecasting at IRI. *Bull. Am. Meteorol. Soc.* **84**, 1783–1796.

Bell, M. J., Forbes, R. M. and Hines, A. 2000. Assessment of the FOAM global data assimilation system for real-time operational ocean forecasting. *J. Mar. Syst.* **25**, 1–22.

Bell, M. J., Martin, M. J. and Nichols, N. K. 2004. Assimilation of data into an ocean model with systematic errors near the equator. *Q. J. R. Meteorol. Soc.* **130**, 873–893.

Davey, M. K., Huddleston, M., Sperber, K. R., Braconnot, P., Bryan, F. and co-authors 2002. STOIC: a study of coupled model climatology and variability in tropical ocean regions. *Climate Dyn.* **18**, 403–420 (D02).

Davies, J. R., Rowell, D. P. and Folland, C. K. 1997. North Atlantic and European seasonal predictability using an ensemble of multi-decadal atmospheric GCM simulations. *Int. J. Climatol.* **7**, 1263–1284.

Gordon, C., Cooper, C., Senior, C. A., Banks, H., Gregory, J. M. and co-authors 2000. The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Climate Dyn.* **16**, 147–168.

Graham, R. J., Evans, A. D. L., Mylne, K. R., Harrison, M. S. J. and Robertson, K. B. 2000. An assessment of seasonal predictability using atmospheric general circulation models. *Q. J. R. Meteorol. Soc.* **126**, 2211–2240.

Huddleston, M. R., Bell, M. J., Martin, M. J. and Nichols, N. K. 2004. Assessment of wind stress errors using bias corrected ocean data assimilation. *Q. J. R. Meteorol. Soc.* **130**, 853–871.

Hurrell, J. W. 1995. Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science* **269**, 676–679.

Jones, P. D., Jónsson, T. and Wheeler, D. 1997. Extension to the North Atlantic Oscillation using early instrumental pressure observations from Gibraltar and southwest Iceland. *Int. J. Climatol.* **17**, 1433–1450.

Levitus, S. and Boyer, T. P. 1994. *World Ocean Atlas 1994, Vol. 4: Temperature*. NOAA/NESDIS E/OC21, US Department of Commerce, Washington, DC, 117 pp.

Levitus, S., Burgett, R. and Boyer, T. P. 1994. *World Ocean Atlas 1994, Vol. 3: Salinity*. NOAA/NESDIS E/OC21, US Department of Commerce, Washington, DC, 99 pp.

Murphy, A. H. 1973. A new vector partition of the probability score. *J. Appl. Meteorol.* **12**, 595–600.

Palmer, T. N. and Anderson, D. L. T. 1994. The prospects for seasonal forecasting – a review paper. *Q. J. R. Meteorol. Soc.* **120**, 755–793.

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M. and co-authors 2004. Development of a European multi-model ensemble system for seasonal-to-interannual prediction (DEMETER). *Bull. Am. Meteorol. Soc.* **85**, 853–872.

Pope, V. D., Gallani, M. L., Rowntree, P. R. and Stratton, R. A. 2000. The impact of new physical parametrizations in the Hadley Centre climate model: HadAM3. *Climate Dyn.* **16**, 123–146.

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V. and co-authors 2003. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.* **108**(D14), 4407.

Rodwell, M. J. and Folland, C. K., 2002. Atlantic air–sea interaction and seasonal predictability. *Q. J. R. Meteorol. Soc* **128**, 1413–1443.

Rodwell, M. J. and Folland, C. K., 2003. Atlantic air–sea interaction and model validation. *Ann Geophys.* **46**, 47–56 (RF03).

Rodwell, M. J., Rowell, D. P. and Folland, C. K. 1999. Oceanic forcing of the wintertime North Atlantic Oscillation and European climate. *Nature* **398**, 320–323.

Ropelewski, C. F. and Halpert, M. S. 1987. Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Mon. Wea. Rev.* **115**, 1606–1626.

Roulston, M. S. and Smith, L.A. 2003. Combining dynamical and statistical ensembles. *Tellus* **55A**, 16–30.

Stanski, H. R., Wilson, L. J. and Burrows, W. R. 1989. Survey of common verification methods in meteorology. World Weather Watch Technical Report. No. 8, WMO/TD 358, World Meteorological Organization, Geneva, Switzerland.

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

Sutton, R. T. and Hodson, D. L. R. 2003. Influence of the ocean on North Atlantic climate variability 1871–1999. *J. Climate* **16**, 3296–3313.

Toth, Z., Talagrand, O., Candille, G. and Zhu, Y. 2003. Probability and ensemble forecasts. In: *Forecast Verification: A Practioner's Guide in Atmospheric Science* (eds I. T. Jolliffe, and D. B. Stephenson). Wiley, New York, 137–163.

Trenberth, K. E. 1997. The definition of El Niño. *Bull. Am. Meteorol. Soc.* **78**, 2771–2777.

Van den Dool, H. M. and Toth, Z. 1991. Why do forecasts for 'near normal' often fail? *Wea. Forecasting* **6**, 76–85.

Wilks, D. S. 1995. *Statistical Methods in the Atmospheric Sciences*. Academic, New York.

Wilks, D. S. 2001. A skill score based on economic value for probability forecasts. *Meteorol. Appl.* **8**, 209–219.