The GloSea4 Ensemble Prediction System for Seasonal Forecasting


Met Office Hadley Centre, Exeter, United Kingdom

(Manuscript received 9 September 2010, in final form 18 November 2010)

ABSTRACT

Seasonal forecasting systems, and related systems for decadal prediction, are crucial in the development of adaptation strategies to climate change. However, despite important achievements in this area in the last 10 years, significant levels of skill are only generally found over regions strongly connected with the El Niño–Southern Oscillation. With the aim of improving the skill of regional climate predictions in tropical and extratropical regions from intraseasonal to interannual time scales, a new Met Office global seasonal forecasting system (GloSea4) has been developed. This new system has been designed to be flexible and easy to upgrade so it can be fully integrated within the Met Office model development infrastructure. Overall, the analysis here shows an improvement of GloSea4 when compared to its predecessor. However, there are exceptions, such as the increased model biases that contribute to degrade the skill of Niño-3.4 SST forecasts starting in November. Global ENSO teleconnections and Madden–Julian oscillation anomalies are well represented in GloSea4. Remote forcings of the North Atlantic Oscillation by ENSO and the quasi-biennial oscillation are captured albeit the anomalies are weaker than those found in observations. Hindcast length issues and their implications for seasonal forecasting are also discussed.

1. Introduction

Seasonal predictions using comprehensive coupled dynamical models of the atmosphere, oceans, and land surface have now been operational for nearly a decade. These predictions have been shown to exhibit useful skill for certain regions and seasons (Graham et al. 2005) and socioeconomic benefits are derived from them in areas such as food production and health (Morse et al. 2005; Challinor et al. 2005). Furthermore, seasonal and decadal forecast systems are crucial in the development of adaptation strategies to climate change: there is no better way of adapting to climate change tomorrow than adapting to climate variability today.

In general, significant levels of skill in intraseasonal to interannual predictions are only found over regions strongly connected with the El Niño–Southern Oscillation, partly because over midlatitudes a higher fraction of the atmospheric seasonal mean variability is related to internal unpredictable variability than to forcings from boundary conditions (Kumar et al. 2007). Thus, on average, the forecast skill over extratropical regions, such as Europe, will be lower than over tropical regions.

However, there are important processes that give rise to predictability over midlatitudes that are poorly represented and/or initialized in the models and systems currently used for seasonal forecasting; for example: land surface (Koster et al. 2004), sea ice (Balmaseda et al. 2010), or stratospheric processes (Ineson and Scaife 2009). Therefore, there is scope to improve the skill of these forecasts. With this aim, the new Met Office seasonal forecasting system (GloSea4) has been designed to be a flexible, easy-to-upgrade system, fully integrated within the Met Office model development infrastructure to help speed up model improvement.

GloSea4 became the Met Office Hadley Centre (MOHC) operational seasonal forecasting system in September 2009. The main differences with its predecessor are:

- all input data for forecasts (not hindcast) comes from Met Office data assimilation systems;
- hindcasts are run in real-time, in parallel to the forecast, to allow model and system changes to be introduced easily;

Corresponding author address: Alberto Arribas, Met Office Hadley Centre, Exeter, EX1 3PB, United Kingdom.
E-mail: alberto.arribas@metoffice.gov.uk

DOI: 10.1175/2010MWR3615.1
time-varying climate forcings;
forecast simulations are initialized weekly to allow a better sampling of the evolution of the climate system and more frequent forecast updates; and
beside the introduction of a new model, scientific enhancements include new initialization, representation of uncertainties and postprocessing strategies.

The paper is organized as follows: section 2 includes a description of the GloSea4 ensemble prediction system and its links with model development. Section 3 presents the verification results for some of the major modes of global seasonal variability: El Niño–Southern Oscillation (ENSO), Madden–Julian oscillation (MJO), and North Atlantic Oscillation (NAO). Section 4 contains a summary and a description of future developments.

2. Description of GloSea4

The Global Seasonal Forecasting System version 4 (GloSea4) is the fourth version of the Met Office ensemble prediction system for seasonal forecasting. GloSea4 has been built around the latest MOHC coupled climate model: Hadley Centre Global Environment Model version 3 (HadGEM3) version 1.1 (released in May 2009).

a. The model: Description and links with model development

HadGEM3_r1.1 is part of the HadGEM3 family: a group of coupled models run at different resolutions but sharing the same infrastructure and physical and dynamical schemes so results are traceable among them. Traceability is the key issue and all HadGEM3 model versions at different resolutions aim to use exactly the same physical parameterizations, except where there is theoretical justification for differences based on the physical processes resolved.

This modeling framework helps us to understand the relevant mechanisms behind any given physical process and to improve its representation in the model by, for example, allowing us to test cleanly whether improvements are due to increases in model resolution or changes (e.g., adding complexity) to the physical parameterizations.

The coupled HadGEM3_r1.1 model (Hewitt et al. 2010) consists of the following components:

- Atmosphere = Met Office Unified Model (UM; Davies et al. 2005), including the land surface Met Office Surface Exchange Scheme (MOSES; Essery et al. 2003).
- Ocean = Nucleus for European Modeling of the Ocean (NEMO; Madec 2008)
- Sea Ice = Los Alamos sea ice model (CICE; Hunke and Lipscomb 2010)

In GloSea4 the resolution of HadGEM3_r1.1 is N96 (approximately 120-km horizontal resolution in midlatitudes) and 38 vertical levels for the atmosphere. For the ocean the ORCA1 grid (tripolar 1° ocean with $\frac{1}{50}$ refinement between $20^\circ$S and $20^\circ$N) and 42 levels in the vertical are used.

The GloSea4 system and the HadGEM3 family of models are part of a wider strategic shift at the MOHC: while previous MOHC climate models were developed exclusively with centennial prediction in mind and then adapted for seasonal prediction, the first use of the HadGEM3 model is for seasonal forecasting as part of GloSea4. In fact, the development of the GloSea4 system and the HadGEM3 family of models are intrinsically linked and all interim coupled model versions are tested using the GloSea4 framework as well as in long-climate integrations before being released. Finally, the seasonal forecasting system is intended to use the latest version available of the HadGEM3 model at the highest resolution affordable so the operational runs can provide valuable information and feedback for further model improvement.

The aim of this strategy is to speed up model improvement to increase the skill of seasonal to decadal predictions, making them more useful for adaptation to climate change and climate variability.

b. The ensemble prediction system

As with any seasonal prediction system, GloSea4 has two components: the real-time forecast and a companion set of hindcasts, also called historical reforecasts, used for postprocessing (bias correction and calibration) and skill assessment. Forecast and hindcast simulations are computed using exactly the same ensemble prediction system and model but with different initial conditions.

There are some important differences in the way the ensembles are run between GloSea4 and its predecessor: First, a crucial technical aspect of GloSea4 is that hindcast simulations (the hindcast period is 1989 to 2002) are run in real time and in parallel to the forecast. This is in contrast to what is done in many centers where all hindcast simulations are completed before running any real-time forecast (Anderson et al. 2007; Saha et al. 2006). Running the hindcast simulations in real time allows model and system changes to be introduced without the burden of repeating the complete, very expensive hindcast set and facilitates the integration with the model development process. However, this does not imply that system and/or model changes are introduced blindly into GloSea4. All changes are tested previously by using: 1) a reduced hindcast set sampling Northern Hemisphere winter and summer seasons in a 14-yr period; 2) long climate integrations; and 3) short-range weather forecasts.
Second, in order to better capture the evolution of the earth system and generate more up-to-date initial conditions, forecast simulations are run in a weekly cycle with initial conditions created every Monday. This setup, similar to the one pioneered at National Centers for Environmental Prediction (Saha et al. 2006) and also followed at the Centre for Australian Weather and Climate Research (Hudson et al. 2010), allows more frequent forecast updates, which is of great value to users. In total, 14 forecast members and 42 hindcast simulations are completed every week.

Issues around the length of the hindcast period are further discussed in section 2c, but it should be noted that the 14-yr period used in GloSea4 (1989–2002) is: 1) well observed, mostly covered by Tropical Atmosphere Ocean (TAO) array and altimeter data, which makes it an optimal period (Balmaseda and Anderson, 2009); 2) long enough to accurately estimate model drifts for bias correction and also to give a reasonable estimate of skill (Vitart et al. 2007; Cusack and Arribas 2009); and 3) relatively close to the present and therefore similarly affected by low frequency variability and trends.

Obviously, a longer hindcast period would be desirable for those regions where there is a large internal variability or if the forecasts are subdivided into several categories (although then we would need to assume that the impacts of nonstationary observing systems, low-frequency climate variability and climate trends are negligible, which is not the case). However, it is our view, shared by our funders, that the main priority now is to improve forecast skill beyond ENSO forecasting and that there is little point in investing large amounts of computing resources increasing hindcast length in an attempt to determine very precisely the existing (relatively low, especially for extratropical regions) levels of skill. At least, not until models have improved and increased resolution is affordable.

The way the forecast and hindcast suites are run as part of GloSea4 is illustrated in Fig. 1. In the forecast suite, atmosphere, land surface and ocean initial states are calculated daily. Every Monday, these initial conditions are fed into the coupled model and a total of 14 forecast members are run. Also every Monday, a 42-member lagged ensemble is created by pulling together all forecast members available from the previous three weeks. An example of this is shown in Fig. 2. The bias correction of this 42-member forecast ensemble is explained in section 2c.

For consistency with the forecast, initial start dates for the hindcast are spread throughout the month but, for simplicity, fixed calendar dates (1, 9, 17, and 25 of every

![Fig. 1. Schematic representation of how GloSea4 forecast and hindcast are run.](http://journals.ametsoc.org/mwr/article-pdf/139/6/1891/4267812/2010mwr3615_1.pdf)

![Fig. 2. Example of Niño-3.4 SST anomalies for forecasts initialized on 15, 22, and 29 Mar 2010.](http://journals.ametsoc.org/mwr/article-pdf/139/6/1891/4267812/2010mwr3615_1.pdf)
The European Centre for Medium-Range Weather Forecasts (ECMWF) Re-analysis (ERA)-Interim (Dee et al. 2009) reanalysis is used to initialize the atmosphere and land surface in the GloSea4 hindcasts as there is no atmospheric reanalysis available using the HadGEM3 model. In the case of land-surface variables (e.g., soil moisture), an anomaly initialization approach, in which ERA-Interim anomalies are calculated and then added to an HadGEM3 model climatology, is followed to avoid the inconsistencies resulting from the very different land surface models used in the HadGEM3 model and the ERA-Interim reanalysis.

The initialization of the hindcast simulations is as follows:

- The atmosphere and land surface components are initialized by interpolating the Met Office operational numerical weather prediction (NWP) analysis from its original resolution of 25 km in midlatitudes (N51L70 as in April 2010) to the model resolution used in GloSea4 (N96L38);

- The ocean is initialized from the GloSea4 Ocean Data Assimilation scheme (ODA) which consists of a parallel version of the Met Office optimal interpolation scheme used for short-range ocean forecasting (Martin et al. 2007). In GloSea4, the ODA runs on a daily cycle, forcing the ocean with the Met Office NWP 6-h analysis fluxes and assimilating ocean SST (in situ and satellite) and ocean profiles (temperature and salinity). Only the ocean initial conditions generated every Monday are used to integrate the coupled HadGEM3_r1.1 model forward;

- Although a sea ice model is incorporated in HadGEM3_r1.1, allowing sea ice evolution during the model simulation, we are currently not assimilating sea ice concentrations or thickness. Instead, sea ice is initialized from a seasonally varying model climatology. The assimilation of sea ice concentration is currently under development and we expect to implement it shortly.

The initialization of the hindcast simulations is as follows:

- The initialization of the forecast is as follows:
  - The atmosphere and land surface components are initialized by interpolating the Met Office operational numerical weather prediction (NWP) analysis from its original resolution of 25 km in midlatitudes (N51L70 as in April 2010) to the model resolution used in GloSea4 (N96L38);

- The ocean is initialized from the GloSea4 Ocean Data Assimilation scheme (ODA) which consists of a parallel version of the Met Office optimal interpolation scheme used for short-range ocean forecasting (Martin et al. 2007). In GloSea4, the ODA runs on a daily cycle, forcing the ocean with the Met Office NWP 6-h analysis fluxes and assimilating ocean SST (in situ and satellite) and ocean profiles (temperature and salinity). Only the ocean initial conditions generated every Monday are used to integrate the coupled HadGEM3_r1.1 model forward;

- Although a sea ice model is incorporated in HadGEM3_r1.1, allowing sea ice evolution during the model simulation, we are currently not assimilating sea ice concentrations or thickness. Instead, sea ice is initialized from a seasonally varying model climatology. The assimilation of sea ice concentration is currently under development and we expect to implement it shortly.

The initialization and the representation of uncertainties (for both initial conditions and model) in the ensemble prediction system have changed markedly from GloSea3.

The initialization of the hindcast simulations is as follows:

- The initialization of the hindcast simulations is as follows:
  - The ocean is initialized in the same way as in the forecast suite but the Met Office NWP fluxes are replaced by ERA-Interim daily fluxes. All subsurface data (profiles, buoys) are taken from the EN3 database (Ingleby and Huddleston 2007). Sea surface data are taken from Pathfinder satellite dataset (Kears et al. 2000; Kilpatrick et al. 2001) and the International Comprehensive Ocean–Atmosphere Data Set (ICOADS) in situ dataset (Worley et al. 2005; Rayner et al. 2006);

Climate forcings (e.g., aerosols, methane, CO2, etc.) are set to observed values up to 2000 and follow the Intergovernmental Panel on Climate Change (IPCC) emissions scenario A1B after that. Ozone is fixed to observed climatological values and includes a seasonal cycle. Observed year-to-year solar variability is specified for the hindcast. For the forecast a constant cycle with an 11-yr period is used. Volcanic aerosol is another factor that can impact the radiative forcing of the atmosphere. There is currently no scheme to explicitly represent this in GloSea4 but, given that the net effect of tropical volcanic eruptions can be felt during approximately two years after the eruptions (Robock and Mao 1995), a simplified scheme (Marshall et al. 2009) has been developed and will be implemented in GloSea4 in the near future. Such a scheme could be switched on in the event of a volcanic eruption in the tropics but it is not suitable for volcanic eruptions outside the tropics.

2) REPRESENTATION OF UNCERTAINTIES IN THE INITIAL CONDITIONS

GloSea4 does not contain an explicit representation of uncertainties in the initial conditions. Instead, a lagged-average approach is followed, with a 42-member forecast ensemble created once a week by pulling together all forecasts available from the last three start dates (see Fig. 2). Thus all simulations initialized on a particular start date have exactly the same initial conditions and differ only because of the stochastic physics schemes used to represent model uncertainties.

There are three main reasons for choosing this approach over what was done previously in GloSea3 where sea surface temperatures and wind stress were perturbed from a central analysis:

- Adding perturbations to a central analysis degrades the initial conditions and the skill of the forecast (Bowler et al. 2008b).

- Similar spread is obtained in the current system with fewer arbitrary choices to define the perturbations (see section 3).

- By following a lagged-averaged approach technique we can have a fully parallel run to the high-resolution atmosphere-only and ocean-only deterministic global
models and the Met Office’s short- and medium-range ensembles. This is extremely useful as it allows us to compare the evolution of model drifts and biases.

3) REPRESENTATION OF MODEL UNCERTAINTIES

Model uncertainties can be addressed through the use of multiple models [e.g., projects such as EUROSIP; Asian Pacific Economic Cooperation Climate Centre (APCC)]. A drawback of these methods is that ensemble members are systematically different from each other and tend to cluster by model (Alhamed et al. 2002) so that the multi-model forecast distribution is not the true probability distribution function of the forecast outcomes.

In contrast, in GloSea4 we have developed a seasonal forecasting system that aims to address model error within a single model environment by using stochastic physics in the model. In general, the spread in GloSea4 has improved with respect to GloSea3 and is comparable to other centers around the world (see section 3). Two schemes, both used in the operational Met Office short- and medium-range ensemble prediction systems (Bowler et al. 2009), are implemented in GloSea4:

- The random parameters scheme (Bowler et al. 2008a) representing the structural uncertainties caused by the choice of empirical-adjustable parameters in physical parameterizations (with a selected group of parameters in the convection, large-scale precipitation, boundary layer, and gravity wave drag parameterizations treated as stochastic variables);
- The stochastic kinetic energy backscatter scheme version 2.0 (Shutts 2005; Tennant et al. 2011) representing the subgrid-scale uncertainty arising from excessive energy dissipated in the semi-Lagrangian advection and numerical schemes.

The growth of spread during the first 80 days is illustrated in Fig. 3. As can be seen, the spread of the system during the first two weeks is small but it grows rapidly afterward, saturating after approximately 45 days. This makes the system suitable for seasonal forecasting but not for monthly forecasting. However, as a future upgrade, it is intended to initialize all forecast runs daily in order to increase the ensemble spread during the first weeks.

c. Forecast bias correction

For bias correction purposes it is necessary to estimate a start-date-and-lead-time-dependant model climatology (Stockdale et al. 1997) with enough hindcast years and ensemble members to avoid sampling issues (Cusack and Arribas 2009).

Data from the Met Office contribution to the ENSEMBLES project (Weisheimer et al. 2009) have been used to assess whether 14 years in the hindcast are sufficient to produce a reasonable model climatology. The ENSEMBLES database contains 46 years of data, from 1960 to 2005, for the same four start dates (February, May, August, and September) and number of ensemble members (nine) used in this paper to evaluate GloSea4.

Results from this analysis are shown in Fig. 4 for the Niño-3.4 and northern European (as described in Giorgi and Francisco 2000) regions at different lead times. The mean $\mu$ and standard deviation $\sigma$ for the 46-yr period were calculated for each lead month. The 46 years were then randomly resampled to produce 40-, 35-, 30-, 25-, 20-, 15-, and 10-yr climatologies. Each random resampling was used to calculate a mean, and the resampling repeated 1000 times. A mean $\mu_n$ (asterisks in Fig. 4) and standard deviation $\sigma_n$ (dashed line shows $2\sigma_n$ in Fig. 4) of the subsample means was then calculated.

As shown in Fig. 4a for Niño-3.4 (one-month lead time), in the case of tropical regions the standard deviation of the random resamples ($\sigma_n$) are much smaller than $\sigma$, the standard deviation of the 46-yr data, even for the 10-yr case. In other words, the sampling error is smaller than the year-to-year variability, which is the signal we are trying to forecast. Similar results were found for other tropical regions and lead times (not shown). Thus, a 14-yr hindcast is sufficient for bias-correction purposes in tropical regions as the associated sampling errors are small.

Results are more complicated for extratropical regions (where there is higher internal variability) as shown for northern Europe in Fig. 4b. An important aspect of this is the existence of climate trends and low-frequency variability. This can be clearly observed for the northern European region (six-months lead time shown) where mean model climatologies calculated only using the more recent years (diamonds) are considerably warmer than model climatologies calculated using data from the whole 46-yr period. Differences are typically of the order of $2\sigma_n$ or bigger (i.e., significant and
not random as corroborated by observational studies using longer datasets; Obregon et al. 2008; Kennedy and Parker 2010). Therefore, under the presence of strong climate trends and/or low frequency variability, using a longer hindcast period does not necessarily imply a better estimation of the model climatology. For example, the use of the 1960–2005 hindcast period would introduce a warm bias in the seasonal forecasts over northern Europe.

Thus, even without taking into account the non-stationarity of the observing system, the issue is less, or at least not only, about sampling error than about the existing climate trends (especially over extratropical regions). The choice of the length of the hindcast is therefore not as simple as the longer the better. The hindcast length must be determined carefully and it will depend on the system’s requirements. In our case, the strategic drive is to improve the skill of the forecast by integrating the new system within the model development process and that is better served by having a 14-yr hindcast (which, as shown, is long enough for bias correction purposes) that can be run in real time, parallel to the forecast.

The process for bias correction in GloSea4 is as follows: every forecast is bias-corrected using the four closest start dates in the hindcast (three members for each start date in a 14-yr period). However, because forecast simulations are initialized on Monday each week while the hindcasts are initialized on fixed calendar dates (1, 9, 17, and 25 of every month), hindcasts are weighted in order to avoid issues derived from simulations starting at slightly different points in the seasonal cycle.

The weighting function chosen assigns a normalized set of weights proportional to

\[ w = e^{-d^2/100} \]

where \( d \) is the lag–lead in days between the forecast and the hindcast start dates.

Different weighting strategies were tested and the sensitivity found was small so the simple function described above was chosen. Examples of the resulting weights are shown in Fig. 5 for (i) a forecast date coinciding with one of the hindcast dates and (ii) a forecast date midway between two hindcast dates. Each week’s 14-member forecasts are bias-corrected according to

\[ F' = F - \sum w_i H_i, \]

where \( i = 1, 4 \) are the four closest hindcast dates to the forecast start dates. Once bias-corrected, all forecasts available from the last three weeks are pulled together to form a 42-member lagged ensemble forecast.

3. System performance and evaluation

The latest skill scores and forecasts are freely available online through the Met Office (see http://www.metoffice.gov.uk/science/specialist/long-range/). In this paper, the evaluation of GloSea4 was completed using a hindcast dataset for the 1989–2002 period and the following start dates:

- 25 January, 1 February, and 9 February (“February start date”)
- 25 April, 1 May, and 9 May (“May start date”)
- 25 July, 1 August, and 9 August (“August start date”)
- 25 October, 1 November, and 9 November (“November start date”)

There are three ensemble members for each individual start date (i.e., nine-member ensemble for the
“February,” “May,” “August,” and “November” start dates). Each simulation is seven months long.

To assess the system performance, relative operating characteristics (ROC) and Brier skill scores (BSS) were calculated for sea surface temperature over the Niño and the Enhanced Ocean Data Assimilation and Climate Prediction (ENACT; ECMWF 2005) regions, and for 1.5-m temperature and precipitation over Giorgi regions (Giorgi and Francisco 2000). The ENACT regions for the Southern Ocean were cut off at 55°S to filter out sea ice, and the original Giorgi region covering Australia was divided into two. ROC and BSS over each area are calculated by aggregating data from each grid point with a 9-grid-box smoothing applied to the probability fields to reduce sampling errors. To minimize sampling issues when defining tercile boundaries, an underlying theoretical distribution is assumed (e.g., normal for 2-m temperature, gamma for precipitation) and fitted to the hindcast data available as described in Cusack and Arribas (2009). The same period, 1989–2002, is used for model and observations. Scores for sea surface temperature were verified against the Hadley Centre Global Sea Ice and Coverage and Sea Surface Temperature (HadISST) dataset (Rayner et al. 2003), for 1.5-m temperature against ERA-Interim (Dee et al. 2009) and for precipitation against the Global Precipitation Climatology Project (GPCP) dataset (Adler et al. 2003).

The color bars on the figures that follow are chosen such that orange-red colors correspond to positive skill relative to climatology (>0.5 for ROC scores and >0.0 for BSS).

Figure 6 shows ROC scores for sea surface temperature for JJA (June–August; forecast initialized in May) and DJF (December–February; forecast initialized in November); that is, forecast months 2–4, for the upper, middle, and lower terciles (all skill plots are for these categories). All regions show positive skill relative to climatology (>0.5). In general, GloSea4 shows improved positive skill over GloSea3 with the exception of Niño regions for the lower tercile from the November start date. The northeast Atlantic region shows positive skill in both seasons in the outer terciles.

Figure 7 shows ROC scores for 1.5-m temperature over Giorgi land regions. Many tropical and subtropical regions show positive skill relative to climatology; however, some extratropical regions show less skill than climatology, notably the middle tercile for JJA forecasts initialized in May and the below-average tercile for DJF forecasts initialized in November. It is worth noting that seasonal predictability is generally higher in the presence of strong forcing (e.g., large ENSO anomalies), which means that the skill in the outer terciles is generally higher than in the middle tercile. In JJA, GloSea4 shows improvements relative to GloSea3 in many regions in the outer terciles, and hardly any regions where skill is degraded. GloSea4 performs better in DJF in the upper tercile, but shows some degradation in some regions in the lower tercile. In both seasons the performance in the middle tercile is very mixed with almost equal numbers of regions showing degradation and improvement.

Figure 8 shows ROC scores for precipitation. Precipitation is known to be a difficult field to forecast and,
encouragingly, the model shows skill relative to climatology in a number of regions (the Maritime Continent, north Australia, the Amazon, India) for the upper and lower terciles. Relative to GloSea3, GloSea4 shows improvement in the outer terciles in JJA in several regions. There are some regions where performance is degraded, particularly in the lower tercile in DJF. In general, skill for most land regions, notably northern Europe, in all terciles is still low.

Figure 9 shows Brier skill scores for Niño and ENACT ocean regions. Brier scores can be decomposed into reliability and resolution (Wilks 1995). Since climatology is perfectly reliable (albeit having zero resolution) even a very good forecast will have lower reliability than climatology. Thus, to give a positive Brier skill score, a forecast needs to provide good resolution, which makes the Brier skill score a very demanding score for seasonal forecasts. It is for this reason that performance over Niño regions is typically reported in terms of deterministic scores such as anomaly correlation coefficients (ACC) and root-mean-square error (see section 3a). As can be seen, even for sea surface temperature, GloSea4 does not produce particularly high BSS. The BSS for 1.5-m temperature over land and for precipitation (not shown) are mostly negative. Forecasts of summer temperatures and winter precipitation over the Amazon show positive BSS in all three categories, but forecasts for other regions are less skillful than climatology in at least one of the categories in each season. Similarly to what has been shown for ROC scores, GloSea4 skill is higher than GloSea3 over extratropical regions, particularly the North Atlantic, but worse for the lower tercile.
over Niño regions for the DJF period. The skill of GloSea4 over the Niño regions is further investigated in the following section.

a. ENSO

Figure 10 shows the forecast plumes for the Niño-3.4 region for forecasts initialized in February, May, August, and December. For most individual years and start dates the forecast model reproduces observations quite well. There are some discrepancies; for example, hindcast initialized in November for 1989–93 tend to be colder than observations and the hindcast initialized in August 1997 overestimate the magnitude of that year’s El Niño event.

The solid lines in Fig. 11 show the centered anomaly correlations between the ensemble mean forecast and observations for the Niño-3.4 region and the four different start dates as a function of lead time. GloSea4 values are high but not significantly different to those of GloSea3 (dotted lines show the 90% confidence interval) for any start date or lead time. The largest differences are found for the November start date at lead times of four and five months (i.e., in March and April, coinciding with the so-called “spring predictability barrier”) when GloSea4 anomaly correlation values are lower than those of GloSea3. This behavior could be related to model biases in the version of HadGEM3 used in GloSea4 during the boreal winter (see root-mean-square error of the ensemble mean in Fig. 12 and mean DJF biases in Fig. 13) as it is plausible that nonlinear interactions between forecast signal and model bias are affecting the hindcast simulations initialized in November.

Figure 12 also shows the spread as a function of lead time for the Niño-3.4 region for each of the four start months. The spread-error relationship for GloSea4 is...
quite good—and improved over GloSea3—for hindcast initialized in February and May, as the ensemble mean error is approaching the ensemble spread as it should be in an ideal ensemble prediction system. However, hindcast initialized in August and November are still underdispersive, a problem that affects most seasonal forecasting systems around the world (see, e.g., Anderson et al. 2007; Hudson et al. 2010). Largest differences between root-mean-square error and spread are found in hindcast from the November start date mainly due to the larger model biases at that time of the year (see DJF mean biases in Fig. 13).

Overall, the relationship between ensemble mean error and spread has improved in GloSea4 compared to its predecessor, showing that a combination of a lagged-average approach with stochastic physics schemes is a viable choice for seasonal forecasting.

b. ENSO teleconnections

Teleconnections between ENSO states and precipitation/temperature patterns across the globe were investigated (for JJA forecasts initialized in May and DJF forecasts initialized in November) by differentiating El Niño years from La Niña years. The following criteria were used to classify years into El Niño–La Niña: anomalies exceeding ±1.0°C for running 3-month averages for the Niño-3.4 region were used. The same set of years is used for the forecast even though, as previously shown, the forecast system is not perfect. Other choices of years gave rise to very similar, nearly indistinguishable, patterns for both the JJA and DJF precipitation forecasts.

Figure 14 shows the average precipitation differences from the model (left) and observations (right), for JJA (top) and DJF (bottom). Overall, the model captures the
difference between El Niño and La Niña years remarkably well. The known teleconnection to the Indian monsoon is well captured as are the dry conditions over the northeast region in Australia. In particular, rainfall patterns over Africa seem to be well represented. The DJF pattern of lower rainfall in southern Africa, with a wetter band across Lake Victoria is well reproduced by the model. Figure 15 shows ROC scores over the Greater Horn of Africa region for precipitation in El Niño–La Niña–only years and over all years. As can be seen, forecast skill is higher in El Niño–La Niña years, confirming the system’s ability to capture ENSO teleconnections over Africa.

c. The Madden–Julian oscillation

The Madden–Julian oscillation (Madden and Julian 2005) is the dominant component of intraseasonal variability in the tropics and in recent years the role of the MJO in the global climate system has been recognized. However, global climate models show serious deficiencies in their attempts to represent it correctly. To allow consistent assessment of MJO in climate models, the U.S. Climate Variability and Predictability (CLIVAR) MJO working group have developed a set of diagnostics (CLIVAR Madden–Julian Oscillation Working Group 2009) based on a combined empirical orthogonal function (CEOFS) analysis of National Oceanographic and Atmospheric Administration (NOAA) outgoing longwave radiation (OLR) and 850- and 200-hPa zonal winds from National Centers for Environmental Prediction (NCEP)–National Center for Atmospheric Research (NCAR) reanalyses over the latitude band from 15°S to 15°N. The model output is projected onto the first two observed CEOFSs.
and the resulting time coefficients are named as real-time multivariate MJO indices (RMM1 and RMM2; Wheeler and Hendon 2004). The RMMs are useful tools in analyzing the MJO in forecasts that extract the low frequency evolution of the MJO.

Based on the RMMs, composite evolution of OLR and 850-hPa meridional winds during different phases of MJO in GloSea4 hindcasts are computed and shown in Fig. 16 along with those in the observations. Each phase in the composite life cycle is approximately six days apart. The origin of convective signals in the Indian Ocean and its eastward propagation are captured remarkably well in the OLR and wind anomaly composites despite slightly weaker amplitudes. The biases are comparatively larger in phases 5–7 over the Maritime Continent and western Pacific. The convergence region east of the maximum convection is also represented well.

**FIG. 10.** Forecast plumes for Niño-3.4 region compared to HadISST and ERA-Interim. Plumes are shown for 6 months, for 4 sets of initialization dates through the year. Cross validation has been used to bias correct these simulations.
GloSea4 prediction skill is evaluated in Fig. 17, which shows the correlations between the GloSea4 RMMs and the observed RMMs for the first 25 days. The forecasts are considered to be skillful when the correlations are greater than 0.5. The forecasts are skillful up to 15 to 16 days for both RMMs, which is well above the skill given by the persistence of initial conditions. Note however that the RMM2 is slightly less skillful than

FIG. 11. Centered anomaly correlation for SST in Niño-3.4 region. GloSea3 is shown in black and GloSea4 in gray. Observations are taken from HadISST. Dotted lines indicate the 90% confidence intervals calculated using Fisher’s $z$ transformation (Snedecor and Cochran 1989).

FIG. 12. Rmse (solid line) and spread (dashed line) for GloSea3 (black) and GloSea4 (gray) for Niño-3.4. Observations are taken from HadISST.
RMM1 at early lead times. This is related to model biases in representing the MJO signal in the west Pacific (Fig. 16).

d. The North Atlantic Oscillation

The NAO is the single largest contributing pattern to European interannual variability; prediction of the NAO is therefore a key requirement for skillful predictions of European winter climate. However, seasonal predictability of the NAO is low due, in part, to the weak link between the extratropical ocean and overlying atmosphere in coupled climate models with resolutions of approximately 100 km and poorly resolved stratosphere (Kushnir et al. 2002). Nevertheless, some links with underlying Atlantic sea surface conditions have been identified in simulations with specified ocean conditions (e.g., Rodwell et al. 1999; Mehta et al. 2000) and recent multimodel intercomparisons show positive, albeit weak, correlation with the observed NAO when SST and climate forcings are imposed in models (Scaife et al. 2008). Furthermore, recent high-resolution studies suggest there is potential to improve the representation of the extratropical ocean–atmosphere coupling in models (e.g., Minobe et al. 2008).

GloSea4 has only small seasonal mean biases in the NAO (Fig. 18) and the predicted NAO mean is within 2 or 3 hPa of the observed value. However, like many seasonal forecast systems, the time series of winter mean

FIG. 13. The SST mean bias over the 1989–2002 period for (top) JJA and (bottom) DJF hindcast climatology [model − HadISST observations; the latter regridded to same resolution as model].
predicted NAO values show only a low correlation (~0.2) with the observed NAO for forecasts initialized around 1 November (25 October, 1 November, and 9 November start dates are used here) and this is not statistically significant. The sharp transition between high and low index NAO that occurred in 1995 is not successfully captured, nor is the run of strongly positive NAO years in the early 1990s, which has been connected

![Figure 14](image-url)  
**FIG. 14.** Precipitation difference (mm day$^{-1}$) between El Niño and La Niña years, for model (left) 1989–2002 and (right) GPCP observations averaged over 1979–2007, for (top) JJA and (bottom) DJF. Yellow–red indicates drier conditions in El Niño years. Note that a longer period was used for the observed climatology as although the overall signal for observed differences for 1989–2002 was broadly similar, it was very noisy. During 1979–2007, for JJA, Niño years are 1987, 1991, and 1997 and the Niña year is 1988. For DJF, Niño years (as of January) are 1983, 1987, 1992, 1995, 1998, and 2003, and Niña years are 1989, 1999, and 2000. Years lying outside the 1989–2002 period are omitted in calculating model field differences.

![Figure 15](image-url)  
**FIG. 15.** The ROC curves for the Greater Horn of Africa for (left) Niño–Niña years and (right) all years.
to volcanic forcing following the eruption of Mount Pinatubo in 1991 (e.g., Stenchikov et al. 2006).

Recent studies suggest that while overall predictability of the NAO may be low, predictable signals arise from remote forcing due to the El Niño–Southern Oscillation (Brönnimann 2007; Ineson and Scaife 2009) and the stratospheric quasi-biennial oscillation (QBO; Boer and Hamilton 2008; Marshall and Scaife 2009). We have already seen that GloSea4 shows high skill in predicting ENSO and the QBO can persist in winter providing it is

Fig. 16. Composite anomalies of OLR (shades) and 850-hPa zonal wind (contours) as a function of MJO phases; (left) the GloSea4 composites and (right) the observed composites. Contour intervals are 0.5.
accurately initialized. However, the connections between these sources of predictability and the extratropics are poorly represented in the low stratospheric resolution model used here and are therefore unlikely to be accurately represented beyond their persistence from initial conditions. Figure 19 shows the composite ENSO and QBO signals from the hindcast set and the corresponding observed anomalies. Overall, the patterns seen in observations are quite well represented in GloSea4—which is encouraging given the low skill of seasonal forecasts in this region—however, the magnitude of the anomalies is weaker. The weakness of these signals is perhaps not too surprising given the vertical resolution of the model used in GloSea4 and the lead time of four months by the end of the winter period. Future upgrades to the model are planned to incorporate improved vertical resolution and a model that simulates the QBO using parameterized nonorographic gravity wave drag (Scaife et al. 2002) in the hope of better capturing these effects and the associated predictability in this difficult region.

4. Summary and future developments

GloSea4, the new Met Office seasonal ensemble prediction system, became operational in September 2009. Apart from aiming to improve the skill of seasonal forecasts, there are two high-level objectives behind its development: to facilitate and speed up model improvement and to help adaptation to climate change and climate variability by providing near-term climate predictions of relevant variables for end users.

To achieve these objectives, GloSea4 was designed to be a more flexible system, better integrated with other Met Office systems than its predecessor was. A crucial technical change is the ability to run forecast and hindcast simulations in real-time, which allows us to upgrade the model and system easily. Numerous scientific changes have been made, including the use of a new, higher resolution coupled model, varying climate forcings, and new ensemble generation methods and post-processing strategies.

The length of the hindcast in GloSea4 is 14 years, which is long enough for bias correction and skill assessment purposes. It is also best suited to the aim of linking GloSea4 with model development. As discussed, issues such as climate trends, low-frequency variability, and nonstationarity of the observing system mean that the choice of the length of the hindcast is not a simplistic “the longer the better” response and, therefore, the hindcast length must rather be determined carefully depending on the system’s requirements. We are currently not applying any further postprocessing to GloSea4 forecasts beyond bias correction. A possibility would be to use the skill measures derived from the hindcast to calibrate the forecasts. There are limitations to what can be achieved by calibration—if the relevant process is not represented in the system the calibration will not improve the results; also, the estimation of forecast skill derived from hindcast data is only approximate because of unavoidable differences.

In terms of skill, our analysis shows that, overall, the skill of GloSea4 forecasts has improved, with more regions showing higher than lower skill when compared to GloSea3. Sea surface temperatures in the tropical Pacific are well predicted, particularly from February, May, and August start dates. ENSO teleconnections, MJO patterns, and the influence of ENSO and the QBO on the NAO are all well captured in GloSea4.

Future upgrades to the system already under development include: assimilation of sea ice concentrations, daily initialization of forecast simulations, a scheme to
Acknowledgments. This work was supported by the Joint DECC and Defra Integrated Climate Programme, DECC/Defra (GA01101). The authors wish to thank the numerous Met Office research staff that have contributed to make GloSea4 possible.

REFERENCES


ECMWF, 2005: ENACT final report. [Available online at http://www.ecmwf.int/research/EU_projects/ENACT/Literature/index.html.]


