Clim Dyn (2011) 37:455–471
DOI 10.1007/s00382-010-0947-3

ECMWF seasonal forecast system 3 and its prediction of sea surface temperature

Timothy N. Stockdale · David L. T. Anderson · Magdalena A. Balmaseda · Francisco Doblas-Reyes · Laura Ferranti · Kristian Mogensen · Timothy N. Palmer · Franco Molteni · Frederic Vitart

Received: 22 December 2009 / Accepted: 15 November 2010 / Published online: 12 March 2011

Abstract The latest operational version of the ECMWF seasonal forecasting system is described. It shows noticeably improved skill for sea surface temperature (SST) prediction compared with previous versions, particularly with respect to El Nino related variability. Substantial skill is shown for lead times up to 1 year, although at this range the spread in the ensemble forecast implies a loss of predictability large enough to account for most of the forecast error variance, suggesting only moderate scope for improving long range El Nino forecasts. At shorter ranges, particularly 3–6 months, skill is still substantially below the model-estimated predictability limit. SST forecast skill is higher for more recent periods than earlier ones. Analysis shows that although various factors can affect scores in particular periods, the improvement from 1994 onwards seems to be robust, and is most plausibly due to improvements in the observing system made at that time. The improvement in forecast skill is most evident for 3-month forecasts starting in February, where predictions of NINO3.4 SST from 1994 to present have been almost without fault. It is argued that in situations where the impact of model error is small, the value of improved observational data can be seen most clearly. Significant skill is also shown in the equatorial Indian Ocean, although predictive skill in parts of the tropical Atlantic are relatively poor. SST forecast errors can be especially high in the Southern Ocean.

T. N. Stockdale (✉) · D. L. T. Anderson · M. A. Balmaseda · F. Doblas-Reyes · L. Ferranti · K. Mogensen · T. N. Palmer · F. Molteni · F. Vitart
ECMWF, Shinfield Park, Reading RG2 9AX, UK
e-mail: Tim.Stockdale@ecmwf.int

Present Address:
F. Doblas-Reyes
Institut Catala de Ciencies del Clima (IC3), Doctor Trueta 203, 08005 Barcelona, Spain

1 Introduction

Seasonal prediction using coupled dynamical models has progressed significantly over the last 20 years. Starting from coupled atmosphere–ocean models of intermediate complexity and spanning the tropical domain (Cane et al. 1986), seasonal prediction has become an operational activity using comprehensive coupled ocean–atmosphere models spanning the global domain. ECMWF has been running seasonal forecasts using coupled ocean–atmosphere models since 1997 (Stockdale et al. 1998), and is now running its third generation system.

In this paper we describe this latest system, and compare its performance with earlier systems, focusing on prediction of sea surface temperature (SST). A future paper will examine some aspects of the atmospheric component of the forecasts; Tompkins and Feudale (2010) also provide an analysis of rainfall forecasts from this system. For comparison with the results presented here, other papers that have examined SST forecast skill of seasonal forecast models in recent times are those of Saha et al. (2006), Jin et al. (2008), Luo et al. (2008) and Wang et al. (2009).

The paper is organized as follows. Section 2 describes the seasonal forecast system, used both for real-time forecasts and the "reference" integrations carried out to calibrate and provide skill estimates for the forecast system. Section 3 gives a brief overview of the SST forecast performance. Section 4 discusses the ENSO forecast skill of the system in more detail, and Sect. 5 examines briefly

performance in the tropical Indian and Atlantic oceans. A concluding discussion is given in Sect. 6.

2 ECMWF seasonal forecast system 3

A new seasonal forecast system known as System 3 (S3) was introduced at ECMWF in 2007. The forecast system includes a coupled atmosphere–ocean model, a data assimilation scheme to create initial conditions for the ocean, and a strategy for ensemble generation. An essential strategy for the development of the ECMWF seasonal forecast system is that the atmospheric components come from the operational medium-range prediction system, albeit run at lower resolution. The development of the ECMWF seasonal forecast system is thus entirely consistent with the so-called seamless prediction initiative (WCRP 2005; Palmer et al. 2008; Hurrell et al. 2009).

New forecast systems are introduced operationally only once every few years, because of the work involved, the value of stability for users, and the need to produce adequate hindcast datasets in order to have a good a priori estimate of forecast skill and reliability. ECMWF's first real-time seasonal forecast system, later designated System 1 (S1), ran throughout 1997 and became effectively operational in December 1997 when forecast products started to be disseminated via the web. System 2 (S2) started running in August 2001 and took over as the primary system in 2002. System 3 started running in September 2006 and became the operational version in March 2007. A detailed description of the development of System 3 is given in Anderson et al. (2007).

2.1 The atmosphere model

The atmosphere model used for S3 is cycle 31r1 (Cy31r1) of the ECMWF Integrated Forecast System (IFS). It is the same code as was used for numerical weather prediction in late 2006, but configured at a lower resolution so as to be affordable for seasonal prediction. The horizontal resolution of S3 is T_L159, with a corresponding grid mesh resolution of 1.125° or about 125 km. There are 62 levels in the vertical, extending to ~5 hPa. (This compares to a T_L95/~200 km resolution and 40 levels extending to 10 hPa used in the preceding S2). Important features of the atmosphere model include two-time level semi-lagrangian numerics with a finite element discretization in the vertical, the Rapid Radiation Transfer Model (RRTM) scheme for longwave and a six spectral interval scheme for shortwave radiation, mass-flux convection, prognostic clouds, a boundary layer scheme with an eddy-diffusivity mass-flux framework, the TESSEL tiled surface scheme with six land tiles and a four-level representation of soil, turbulent orographic form drag, and sub-grid scale orographic drag. A comprehensive model of the ocean surface waves and their interaction with the atmosphere is also included. The model numerics permit a 1-h time-step. Details of the Cy31r1 version of the model, including extensive references, are available online at www.ecmwf.int/research/ifsdocs.

2.2 Atmosphere and land surface initial conditions

Initial conditions for the atmosphere model are derived from the ERA-40 analyses for dates up the end of 2001, and from the operational ECMWF analyses thereafter. The high-resolution operational analyses are interpolated to the target T_L159 resolution, using standard software which handles the hybrid vertical coordinate system under the imposed change in orography. Some important fields for seasonal prediction, such as soil moisture and snow depth, can be detrimentally affected by the interpolation. In some cases, problems with interpolation seem inevitable, such as when mapping a mountain range with snow-covered peaks and warm valleys to a low resolution elevated "bulge" that may or may not be high enough to sustain snow cover, and attempting to be consistent in this when the input resolution varies over time.

There are also problems with the input land surface analyses, which vary over time due to the quality of input data (e.g. issues with snow cover during some years of ERA-40), and, for the operational system, changes in the data assimilation system itself. S3 does include useful information on the state of land surface initial conditions, to the benefit of the forecasts. However, for the reasons mentioned above, there remain significant sources of error in the land initial conditions, in some cases large enough to influence the seasonal forecasts, particularly of 2-m temperature.

2.3 Specified forcings for the atmosphere

There are several sources of variation of climate which are not modelled interactively in System 3, but which can be specified. Foremost of these is the time-variation of greenhouse gases in the atmosphere model. Changes in CO_2 have a substantial impact on seasonal forecasts, even at a time range of only a few months, when comparing forecasts and hindcasts made decades apart (Doblas-Reyes et al. 2006), although small changes or inaccuracies in the trace gases compared to the overall trend are unlikely to be relevant. Thus an approximate time-history of CO_2, methane, and CFCs is specified, based on observed values up to 2000 and values derived from the IPCC A1B scenario beyond this (for the scenario values see Appendix II of the IPCC TAR—our CO_2 values follow the reference

calculation of the BERN carbon cycle model). For the hindcasts the year-to year variability in the solar constant is specified, although a fixed value is used after 2000. There is no evidence that solar variability has much impact on the S3 forecasts, although the recent results of Meehl et al. (2008) are noted.

Another source of variability in the observed radiative forcing of the atmosphere is volcanic aerosol. It was decided to switch off the option for time variation of volcanic aerosol in the hindcasts, since real-time volcanic aerosol analyses are not available for the operational forecasts. A sufficiently large volcanic eruption would invalidate the assumptions of our forecast system.

2.4 The ocean model and ocean initial conditions

The ocean model used is HOPE, and is almost identical to the version used in S2. The resolution remains effectively $1° \times 1°$ in mid-latitudes, with a 0.3° meridional resolution at the equator. There are 29 levels in the vertical with the highest resolution (10 m) near the surface. A sophisticated ocean data assimilation system is used to prepare ocean initial conditions for the forecasts. This assimilation system has been substantially developed since S2. Some of the major changes are analysis of salinity as well as temperature, balanced velocity adjustments, an adaptive bias correction scheme (Balmaseda et al. 2005) and use of altimeter data. A full description of the ocean model and the ocean assimilation system can be found in Balmaseda et al. (2008). The ocean model used for the forecasts is not quite identical to the model used in the assimilation system. It was found that for data assimilation, a lower value for the shear-dependent horizontal mixing of heat and salt was beneficial, whereas for the forecasts themselves, early testing showed that a higher value gave significantly better forecasts of equatorial Pacific SST anomalies. The difference in forecast performance was large enough to warrant the use of the higher value of mixing in subsequent development of the system. Note that because of pervasive model error, there is no "right" value which would make the model match reality. Since the ocean model is subject to different forcings and treatment during assimilation and forecasts, the possibility of different "best fits" of a parameter is not unreasonable.

2.5 Treatment of sea-ice

S3, like all ECMWF seasonal forecast systems to date, does not have a physically-based model of sea ice. Instead, for both forecasts and hindcasts, a module within the ocean model specifies persistence for 10 days of the initially specified fractional ice cover, taken from the ECMWF operational NWP analyses or ERA-40 reanalyses, as appropriate. After a forecast time of 10 days, the specified fractional ice cover is a linear combination of the initial sea ice cover and the climatological ice cover valid at the specification date. Beyond a forecast time of 30 days, the specified sea ice cover is simply defined by linear interpolation of climatological monthly mean values. This specification appears to give reasonable behaviour for most of the period of the hindcasts, but is not very appropriate in the Arctic for the last few years (2005 and beyond, say), due to the strong observed trends to reduced ice cover. This is of concern for the real-time forecasts in particular (Balmaseda et al. 2010), and will be addressed in future systems.

2.6 Structure of the forecast and hindcast ensembles

As in S2, the ocean initial conditions in S3 are provided not from a single ocean analysis but from a 5-member ensemble of ocean analyses, created by adding perturbations to the wind forcing used in the analysis. The ocean initial conditions are further perturbed by adding sea surface temperature perturbations to the five member ensemble of ocean analyses (Vialard et al. 2005). The amplitude of the wind perturbations in S3 has been reduced, and more accurately represents the uncertainties in the wind forcing used; the SST perturbations are also reduced (Weisheimer 2005).

The initial atmospheric conditions are perturbed with atmospheric singular vectors, calculated as per the ECMWF medium-range ensemble forecast system for this atmospheric model version (a combination of initial singular vectors, evolved singular vectors, and targeted singular vectors in the tropics). Stochastic physics (stochastic perturbation of physical tendencies with a 6-h decorrelation time scale) is active throughout the forecast period, again as in the medium-range ensemble forecast system (Buizza et al. 1999).

Each real-time ensemble forecast comprises 41 integrations, and the hindcasts consist of 11 member ensembles for the same start month of each of the 25 years 1981–2005, thus creating a calibration probability distribution function of 275 members. Every member of every ensemble has a start date of the first of the month.

The seasonal integrations are generally 7 months long. Once per quarter, 11 members of the ensemble are extended to a length of 13 months, with the intention of allowing an "ENSO outlook" to be produced. A subset of the hindcast integrations are also extended to 13 months, once per quarter, with a 5 member ensemble.

3 An overview of SST forecast performance

Forecasts of SST are bias-corrected for the mean drift in the coupled model forecasts (Stockdale 1997). Examples of

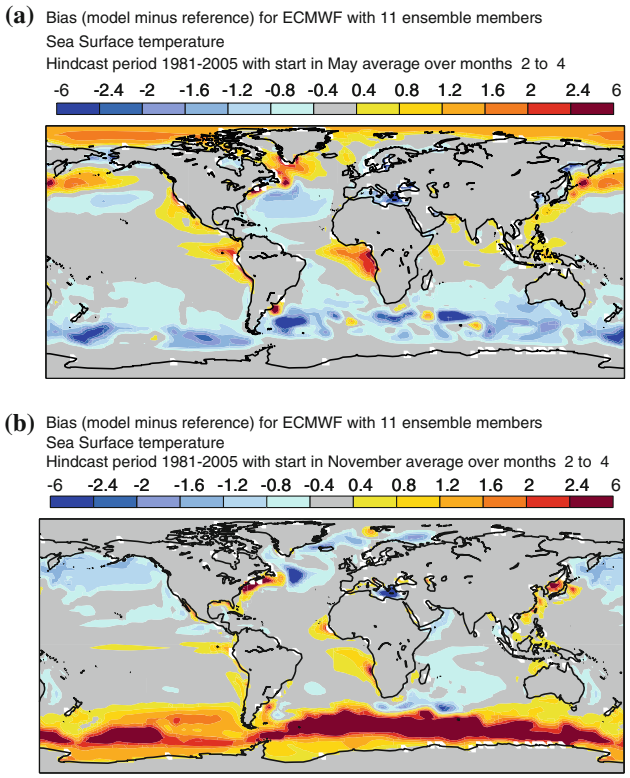

Fig. 1 Mean biases in the 1 month lead S3 seasonal mean SST forecasts: **a** JJA for forecasts started on the 1st May, and **b** DJF for forecasts started on the 1st November. The contour interval is 0.4°C

SST bias are shown in Fig. 1 for June–August (JJA) for forecasts initialized on 1st May and December–February (DJF) forecasts initialized on the 1st November. The bias is seasonally dependent and increases with lead time, but remains small at seasonal timescales in most regions of the ocean, particularly in the tropics. The west Pacific is notably improved compared to previous ECMWF seasonal forecast systems.

An overview of the ability of S3 to predict SST anomalies is given in Fig. 2 which shows the temporal anomaly correlation coefficient (ACC) of 1 month lead seasonal mean forecasts with the verifying analysis, calculated at each grid point for JJA (panel b) and DJF (panel e). The S3 ACC can be compared with that obtained by persistence of the observed monthly mean SST anomaly from the month immediately prior to the start of the forecast (panels a, d). A model-derived estimate of the "perfect model" predictability limit of ACC is also shown (panels c, f), obtained by correlating the ensemble mean forecast against each individual ensemble member, used as a proxy for "truth". This predictability limit represents the score that could be achieved with a perfect model, assuming the initial condition errors were of the size specified by the ensemble initial conditions, and that the real world has predictability characteristics similar to those of the S3 model. The existence of model errors means that the true predictability limit could be higher or lower than the model estimate; in particular, the real world might have higher predictability if there are sources of predictability which are not represented in our model.

Comparison of the panels of Fig. 2 shows that S3 is generally an improvement over anomaly persistence, but still substantially below the model-estimated upper limit of predictability. Exactly where the model performance lies between these two values, in essence defining zero and perfect skill, varies according to both region and time. Anomaly persistence usually loses correlation fairly rapidly, enabling the model to do substantially better than it at longer lead times in a number of regions (not shown); but at these longer lead times the model performance remains substantially worse than the model-estimated upper limit.

Figure 2 demonstrates that the relative advantage of S3 forecasts over persistence for the tropical Pacific is highest for the May starts, although the actual ACC is highest for November starts, and the perfect model estimate suggests high ACC is possible in both seasons. Section 4 gives a detailed analysis of ENSO forecast skill and its seasonal variation.

At mid and high latitudes in both hemispheres, Fig. 2 shows that winter SST is potentially more predictable than summer SST, as to be expected from the enhanced ability of the ocean initial state to influence SST when mixed layers are deep. However, S3 has a significant failing: despite the high levels of potential predictability, the Southern Hemisphere winter SST forecasts have low levels of skill, in many cases below that of persistence. One possible explanation of these results is that the ocean initial states in these latitudes during the 1981–2005 period might contain little useful sub-surface information—perhaps not surprising, given the almost complete lack of in situ ocean data. One would hope that with the ARGO array now in place, SST prediction in this part of the ocean will improve, although this has yet to be confirmed.

In assessing plots such as Fig. 2, it is important to remember that, in almost all applications, predicting SST is just the first stage in predicting user-relevant seasonal climate anomalies. The relatively low level of SST prediction skill in many mid- and high- latitude regions is thus of limited direct concern: what matters is the level of predictability and skill in areas where the atmosphere is most sensitive to SST anomalies. In general the key areas for predicting SST are the tropical oceans, particularly those regions where small local changes in SST can lead to large spatial shifts in atmospheric deep convection. ENSO is the pre-eminent example of this, and will be the focus of much of this paper. Mid-latitude SST prediction is of some interest as a *diagnostic* of prediction of seasonal anomalies in the mid-latitude atmosphere, particularly in summer when mixed layers are shallow. However, assessment of

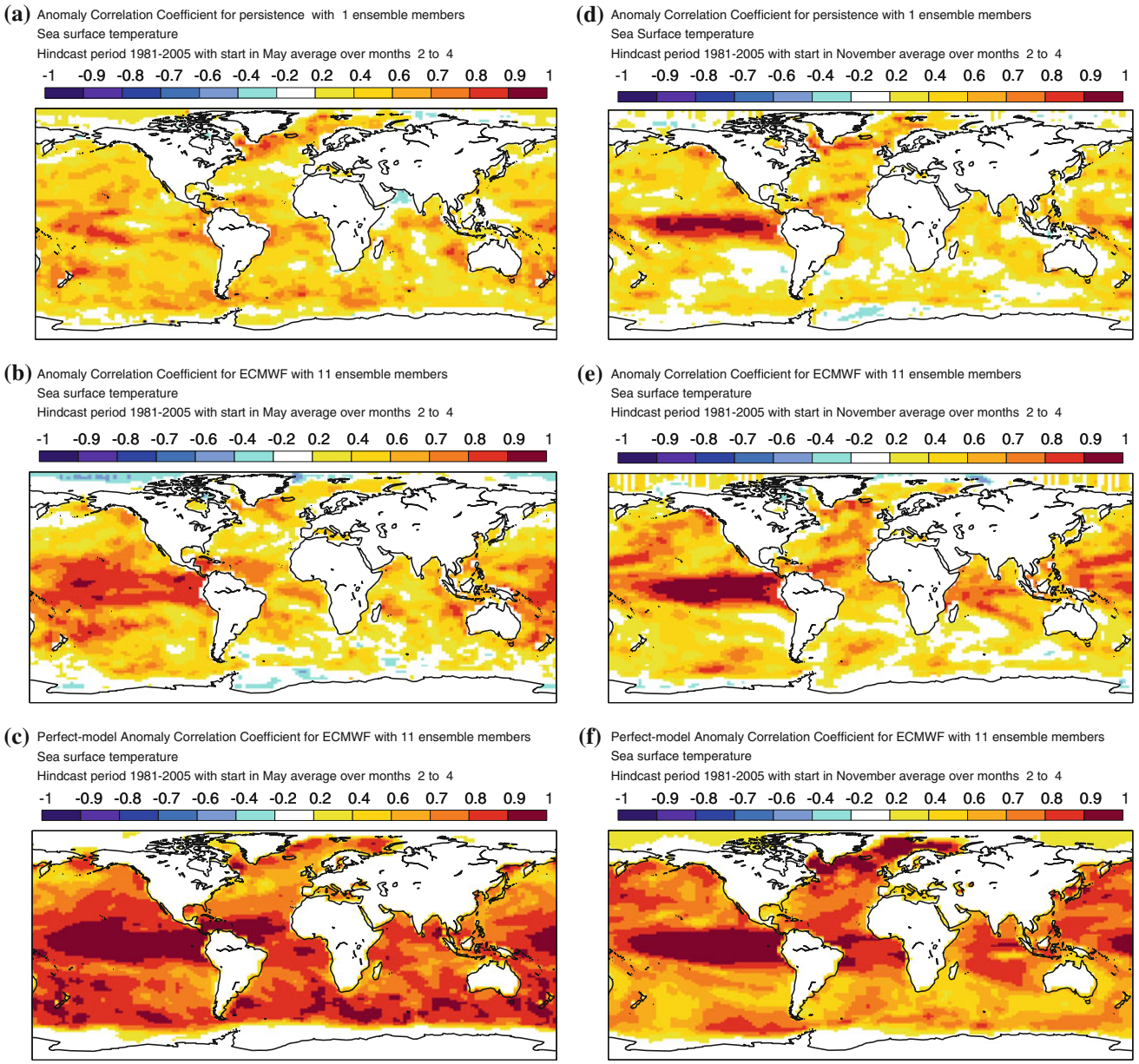

Fig. 2 Spatial map of SST ACC, for anomaly persistence (**a** and **d**), S3 forecasts (**b** and **e**), and the S3 estimated predictability limit (**c** and **f**), for forecasts verifying in JJA (**a**–**c**) and DJF (**d**–**f**). Calculated from hindcasts for 1981–2005. Initialization dates are 1st May and 1st November

seasonal forecast skill in mid-latitudes is bedevilled by low signal-to-noise ratios and limited sample sizes, and skill measures must be used with due caution. We will not consider mid-latitude skill any further in this paper.

4 ENSO forecast skill

4.1 ENSO forecast scores and diagnostics

The largest driver of interannual variability in the troposphere is the variability in equatorial Pacific SST associated with El Nino-Southern Oscillation (ENSO) phenomenon. Assessing the skill of ENSO SST predictions is thus a fundamental requirement for any seasonal forecasting system. Since there are sometimes variations in how scores are calculated, we will be explicit about our methods. Calendar monthly mean SSTs from the model are verified against those of the chosen observational dataset: HadISSTv1 SST (Rayner et al. 2003) up to 1981, and NCEP OIv2 (Reynolds et al. 2002) from January 1982 onwards. These datasets attempt to represent the actual variation of SST from month to month—they are not filtered or designed to represent only certain modes or frequencies of ENSO.

Deterministic scores are based on the ensemble mean of each forecast. The bias-correction of the hindcasts uses cross-validation: for each of the 25 years in the hindcast dataset, the bias is removed using only the other 24 years. Various aggregate statistics are calculated from the bias-corrected forecasts. The most familiar are the root-mean-square error (RMSE) and the temporal anomaly correlation coefficient (ACC). When comparing ENSO-related SST scores, it is generally found that RMSE is more stable than ACC with respect to changes in the period being assessed, although there is still some sensitivity, and comparisons of scores between systems should always be made for an identical set of cases. The RMSE is calculated by averaging the mean square error across all start dates, not by averaging the RMSE for different seasons. Likewise, the ACC is based on considering the full set of anomalies together—the relative amplitude of e.g. summer and winter anomalies will affect the score. Since ENSO SST index forecasts are issued in terms of anomalies relative to 1971–2000, the ACC is calculated using this period to define the mean, rather than the mean of the years being assessed.

A particularly useful statistic is the mean square skill score (MSSS) relative to climatology, defined as $(1 - \mathrm{MSE}/\mathrm{MSE}_{\mathrm{clim}})$, where $\mathrm{MSE}_{\mathrm{clim}}$ is the mean square error of a climatological forecast. The MSSS is in essence just a transformation of the RMSE score, but unlike the RMSE score, it is easy to see at a glance how the forecast compares with a climatological forecast. In the case that the amplitude of the forecast anomalies matches the observed amplitude, the MSSS is simply the square of the ACC. However, unlike the ACC, the MSSS is sensitive to the amplitude of the predicted anomalies: if the model predicts anomalies which are too large or too small, the MSSS will suffer.

One might argue that ACC is more appropriate than MSSS for a system simply designed to predict SST anomalies—the forecast amplitudes can always be rescaled afterwards if necessary, although with a slight loss of skill since the scaling is only estimated. However, in a "single-tier" seasonal forecast system, the atmospheric model reacts directly to SST without any scaling of anomaly. A measure sensitive to the amplitude of SST anomalies is therefore appropriate.

The ACC, RMSE and MSSS are all deterministic scores, applied to the ensemble mean forecast. In contrast, the actual forecasts of SST by S3 are intrinsically probabilistic, with the ensemble spanning a range of outcomes. Although we will not consider probabilistic skill measures here (see Weisheimer et al. 2009; Palmer et al. 2004), we will consider the ensemble spread of the forecasts as an important diagnostic: the spread of the forecast (the unbiased estimator of the standard deviation of the ensemble members about the ensemble mean) would match the RMSE of the ensemble mean forecast in a perfect ensemble forecasting system. The extent to which it does not, gives information on the performance of the system in representing uncertainty in the forecasts.

A final useful diagnostic is the amplitude of the SST anomalies generated by the model, compared to those observed. This is measured by the ratio of the standard deviation of the anomalies of individual model forecasts to the standard deviation of the observed anomalies. Although the ensemble mean forecast will have a standard deviation that becomes small at long lead times, due to averaging over the unpredictable component of the forecasts, the individual ensemble members of a near-perfect forecast system should have an amplitude which is unbiased compared with the observations. The standard deviation of the model is estimated by considering all ensemble members, with the anomalies defined by using a cross-validated bias correction to correct the model forecasts to absolute values, before subtracting an observed long-term climate. The observed standard deviation is defined using observed SST for the same verification dates, and the subtraction of the same long-term climate.

4.2 Comparison of skill with previous systems

Our primary measure of ENSO is the NINO3.4 SST index, defined as monthly mean SST averaged over the area 5N–5S, 120–170W. Figure 3 compares the forecast performance for NINO3.4 of S3 and the two previous operational ECMWF seasonal forecast systems (S2 and S1), for the common period 1987–2002. The RMSE (Fig. 3a) and MSSS (Fig. 3b) show a clear progression of improving skill with each new forecast system, and RMSE is now substantially lower than in the first system. The "time gap" between each system is about 5 years. The 95% confidence interval for the S3 RMSE is shown with thin red lines, and represents the uncertainty due to the limited ensemble size on the given set of 192 cases (the uncertainty is similar for the other systems, but omitted for clarity; the difference in skill level that might be expected in a *different* 16 year period is harder to constrain with a confidence interval—see discussion in Sect. 4.4). Figure 3a also shows, as dashed lines, the ensemble spread of the forecast from each of the systems. S1 had no explicit treatment of the initial condition errors. S2 used initial condition error estimates which were too large in some aspects. S3 has a better estimate of the magnitude of initial condition uncertainty coming from SST and wind stress forcing, although there are other sources of initial condition error which are presently neglected.

Although the ensemble spread in the first month is clearly related to the specification of initial condition uncertainty, by 3 months and beyond the ensemble spread is largely determined by the properties of the forecast

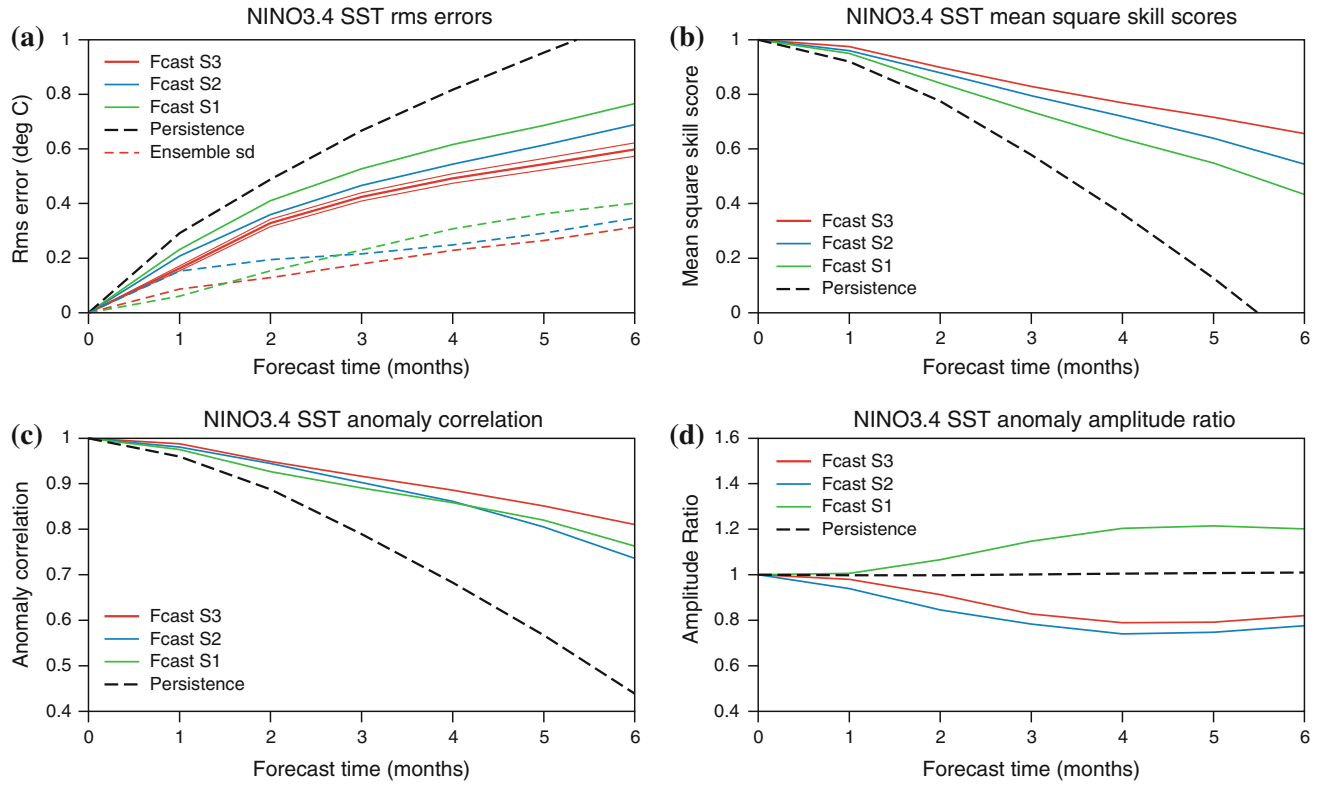

Fig. 3 Comparison of NINO3.4 SST forecast statistics from ECMWF S3 (*red*) with previous ECMWF systems S2 (*blue*) and S1 (*green*), and with anomaly persistence (*black dot-dashed*). **a** rms error; **b** MSSS relative to climate; **c** ACC; **d** amplitude ratio of forecast anomalies to observed anomalies. Calculated from forecasts made every month for the common period 1987–2002, in each case using the first five ensemble members from each system. In **a**, the *thin red lines* give the 95% confidence interval on the S3 rms error for the given period

model (Vialard et al. 2005). Analysis of wind variability in S3 suggests that it is slightly weaker than observed, but as long as only a linear relationship is assumed between wind and SST noise on monthly timescales, the affect of this on SST spread should be modest. Ensemble spread in SST is also sensitive to the mean state of the coupled system (a stronger cold tongue tends to give larger SST anomalies per unit of wind forcing), but the S3 mean state, though not perfect, is not dramatically wrong. Overall, a very rough estimate of the fundamental limit to NINO3.4 predictability might be a standard deviation of order 0.4°C at a 6--month lead time. The aim of this estimate is simply to put the performance of S3 in context—there has been real improvement over the last 10 years, but there is still scope for substantial improvement in the years (and indeed decades) ahead.

The amplitude of NINO3.4 anomalies in S3 is modestly damped, being typically 80% of the observed amplitude at lead times of 4–6 months (Fig. 3d). This is a slight improvement on S2 (which was damped more strongly), and contrasts with S1, which was overactive by about 20%. The differing amplitudes explain the relationship between RMSE and ACC (Fig. 3c)—in particular, that although S2 has a clear advantage over S1 in terms of RMSE, this is largely because of the damping of the anomalies—the improvement in ACC was less significant.

Looking at other regions in the equatorial Pacific (not shown), S2 is a clear improvement over S1 in the NINO3 region in both RMSE and ACC, but S2 is actually worse than S1 in the NINO4 region. This demonstrates a common result when testing new systems—improvements in one aspect of the forecasts are often matched by deterioration elsewhere. In this regard, S3 is an unusually successful system for predicting ENSO-related SSTs—it outperforms S2 at every lead time in every region by every measure under study. In general, it also beats every previous system including versions tested but not made operational, with one interesting exception: S1 remains the best ECMWF operational seasonal forecasting system in terms of ACC in the far west Pacific, for example in EQ3 (5°N−5S, 150E−170°W). The relatively high ACC of S1 at months 3 and beyond comes together with a high amplitude, 120–140% of observed, and significantly higher RMSE than S3. In NINO4 (5°N−5S, 160E-150°W), S1 has essentially the

same ACC as S3, but with an amplitude that is close to 20% too big instead of 20% too small. The high correlation and large amplitude SST anomalies produced by S1 in the west-central Pacific probably accounted for its success in producing robust ENSO forced climate signals around the world.

4.3 Dissecting the skill of S3

Before further examination of the statistics, a presentation of the actual forecasts produced by S3 is given. Figure 4 shows the time evolution of the S3 forecasts of NINO3.4 SST, taking forecasts starting at 3 month intervals, together with the verification. The overall skill in predicting SST anomalies is immediately apparent, with both El Nino and La Nina events often predicted reasonably well, and only a limited number of "false alarms". The ENSO cycle is quite well phase locked to the seasonal cycle during the years covered by these plots (1981-present), and so a seasonal dependence of the forecasts is seen. Forecasts from February initial conditions (panel a) do well at predicting the onset of major El Nino events such as 1982, 1986, 1997 and 2002, although the model sometimes warms too readily, most notably in 1990. The spectacular demises of the 1982/83 and 1997/98 events are also well handled. Forecasts from May initial conditions (panel b) do well at the intensification of the main El Nino events listed above, and also the development of the La Nina events, especially 1995, 1998 and 2007; 1984 and 1988 were both successfully forecast as La Nina conditions, but the details were not correct. Forecasts from August initial conditions attempt to predict the amplitude of El Nino peaks, which they generally do well, although it is noticeable that the spread in the forecasts is wider here than in the ENSO onset phase, and the amplitude of some extremes is missed (e.g. 1982 and 1988, but not 1997 and 1998). Finally, forecasts from November initial conditions generally predict well the decline and termination of El Nino or La Nina conditions—however, damped persistence also works well for forecasts from this time of year. Despite the many successes, it is also clear that there are a number of cases where the forecasts are wrong, differing substantially from the observed evolution of SST.

Considering now statistics of forecast skill, Fig. 5 shows the seasonal variation of NINO3.4 SST scores at 4-month lead, for the period 1994–2007 (panels a and b). The anomaly correlation plot is striking, in that the forecast ACC is always high (0.8–0.9) and almost independent of season, in contrast with the persistence forecast, which shows a strong "predictability barrier" for forecasts made from boreal spring to summer. Although sub-surface information should allow some improvement relative to SST persistence (McPhaden 2003), it is the ability of S3 to overcome this predictability barrier so completely which gives it high levels of overall skill.[1] For forecasts verifying in boreal winter (November–February), the model forecasts give little improvement over persistence, in either RMSE or ACC. Indeed, the forecasts verifying in November–February are quite poor by some measures: RMSE is larger than at other times of year, and substantially larger than the predictability limit estimated by the model (the dashed line in Fig. 5a). Figure 5c, d show the equivalent plots to Fig. 5a, b, except for the longer period 1981–2007. The overall seasonal dependence is very similar, although the RMSE in forecasts verifying in spring and summer is not quite as low as it is in the more recent period. Persistence is also a little less successful in predicting boreal winter SST anomalies. As discussed later, there are reasons for believing that the ENSO SST performance in recent years is more relevant for today's forecasting system than that of the full hindcast period. When looking at the details of the seasonal dependence, though, inadequate sampling is an issue of particular concern for shorter periods. As such, both sets of plots are shown.

A few other points are worth noting about the apparent seasonal dependence of the S3 forecast skill, with the important caveat that some "features" may be artifacts of the limited sampling. For the sake of brevity, figures are not shown. Although there is not much of a predictability barrier in NINO3.4 SST, there is a moderate effect for NINO3 SST anomalies, particularly for the 1994–2007 period, where the ACC drops to around 0.7 for 4-month lead forecasts verifying in May and subsequent months. For NINO3.4, there is a dip in skill for short range forecasts verifying in May and June—indeed the shorter lead forecasts of May/June SST can be worse than the longer lead ones. For persistence forecasts, the key month is July—anomaly persistence forecasts made from 1st July onwards (i.e. using the observed June or later SST anomalies) all have high ACC skill, regardless of lead time.

The seasonal variation in predictability seen in Fig. 5a broadly follows the mean seasonal cycle of oceanic upwelling which peaks in September—strong upwelling goes together with the biggest spread in the ensemble forecasts. The effect first becomes apparent at 2 months' lead, and grows with lead-time. It may be related to the upwelling of anomalies from the ocean subsurface. The growth of the effect with lead-time suggests the spread may be associated less with uncertainty in the initial conditions, and more with the accumulated effect of unpredictable wind variations during the integration.

[1] S2 also had little seasonality in its forecast skill (van Oldenburgh et al. 2005) though it's skill level was lower than S3.

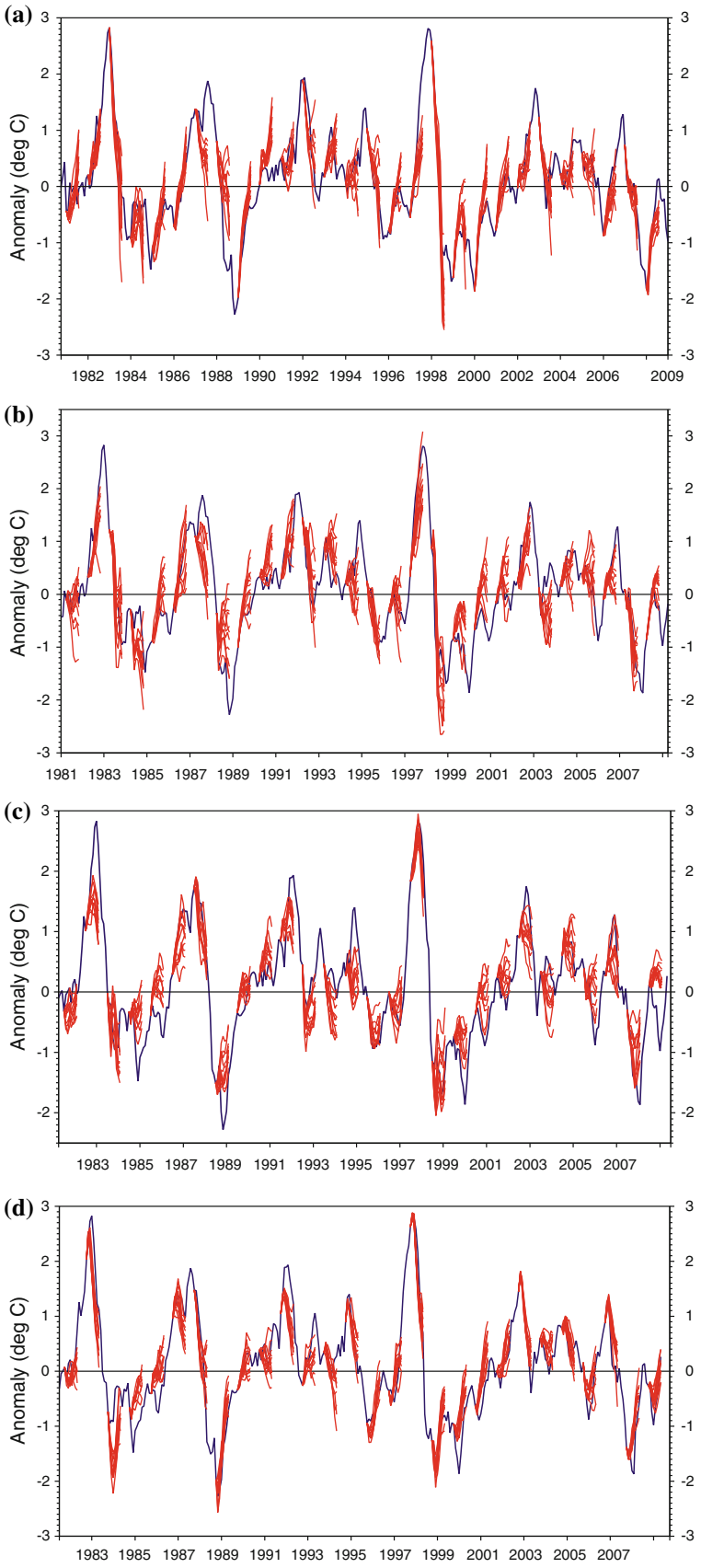

Fig. 4 Forecasts of NINO3.4 SST anomalies from S3, starting from **a** February, **b** May **c** August and **d** November

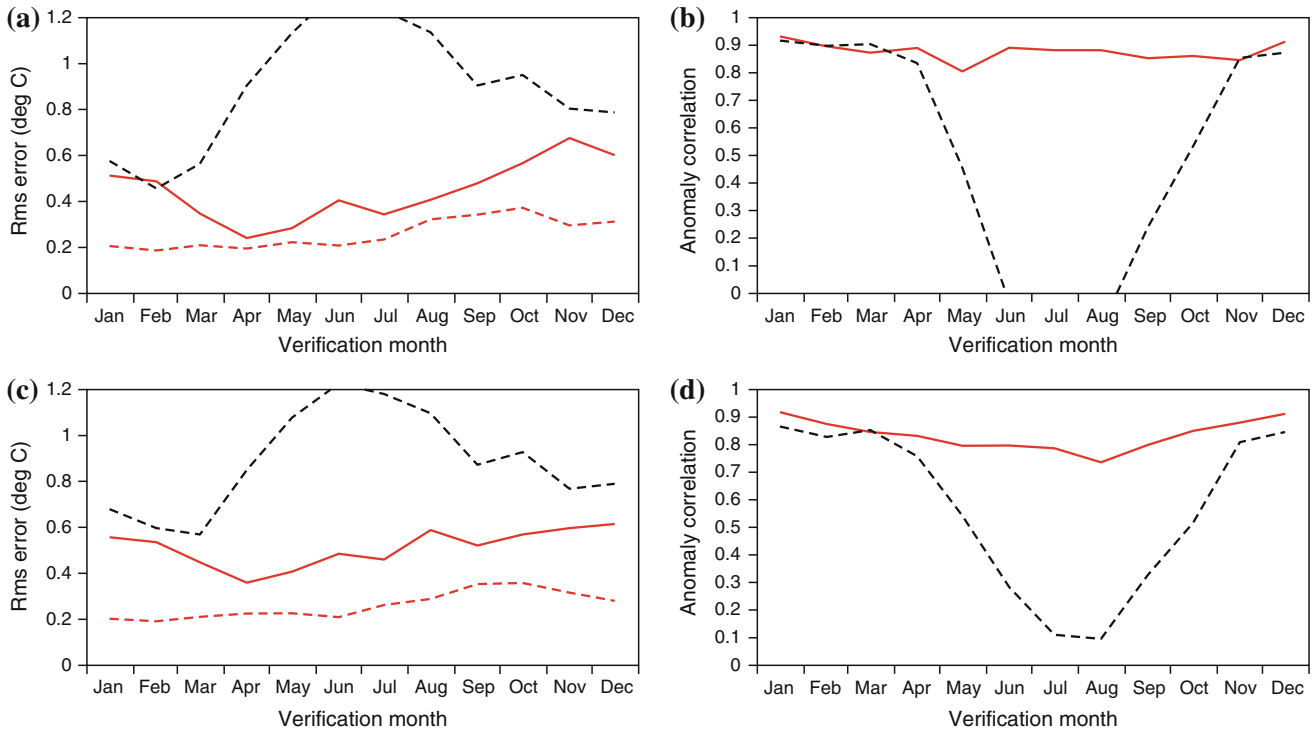

Fig. 5 Seasonal variation of **a** RMSE and **b** ACC for 4 month lead NINO3.4 forecasts from S3, for the years 1994–2007. **c** and **d** the same, but for 1981–2007

4.4 Time evolution of the hindcast skill

The S3 hindcasts and forecasts extend from 1981 to the present day, during which time there has been a considerable evolution of the observing system. Do these improvements in the observing system lead to an improvement in ENSO forecast skill?

There is a serious statistical obstacle to demonstrating improved skill over time. ENSO SST forecast skill, especially at longer range, is very variable. If more recent forecasts have lower errors, can we exclude this being a chance fluctuation? Or, if the errors are not lower, might chance be partially masking improvements in skill? Testing the significance of differences in skill for different time periods is not straightforward. The forecast errors are correlated in time, but more than this, examination of the time-series of errors shows that they cannot reasonably be treated as an AR1 process, which, if this were the case, would allow a simple correction for the degrees of freedom (Zwiers and von Storch 1995). For example, the time series shows a clear relationship between forecast error and ENSO, and since ENSO variability changes over time, we would need to develop an appropriate model for the ENSO activity/forecast error relationship, and use it to remove the ENSO dependence of skill, before we could proceed further with the analysis. Rather than attempt a rigorous analysis of the time evolution of forecast skill, we focus on a few robust statistics which shed light on the change in ENSO forecast skill over time.

The full 1981–2007 verification period can be divided neatly into two approximately equal parts: 1981–1993, and 1994–2007. The dividing line is the point at which the TAO array became operational in the equatorial Pacific, and also coincides very nearly with the start of altimetry. This date defines the biggest single change in the ENSO observing system. The 13 and 14 year periods either side of this date have similar and relatively high levels of ENSO activity. Figure 6a shows the RMSE scores for NINO3.4 for the two periods, for both persistence and for S3.

The NINO3.4 persistence forecasts have similar scores for the two periods, consistent with the similar levels of ENSO variability, and are in fact marginally better in the 1981–1993 period. The S3 ensemble spread for the two periods is also very similar, giving no evidence of any marked differences in predictability. Despite the similarity of the persistence skill, the S3 RMSE is markedly lower in the more recent period. This substantial improvement in RMSE is matched by noticeable improvements in ACC and MSSS. The period with the better observing system has a substantially improved ENSO SST forecast skill. A formal significance test will not be attempted here, because of the difficulties discussed above, but it can be noted that the

difference in skill is large compared to that typically found when comparing different forecast systems over a common 13 year period, and cannot be explained by trivial sampling errors. The results fit with those of Balmaseda and Anderson (2009), who show using observing system experiments that the TAO moorings and altimetry add to the skill of the S3 forecasts.

Similar improved skill in the later period holds for NINO3 SST, but for NINO4 the result is different (not shown). For NINO4 the RMSE of persistence and of S3 are both somewhat higher in the later period, at least for lead times beyond 2 months. The S3 MSSS, which normalizes the errors against the variability, is very similar for both periods, and no improvement in the later period relative to the earlier one is visible. Conceivably, some slight change in predictability characteristics might mask a modest skill improvement. However, NINO4 is the region where we have strongest evidence of model error impacting negatively on forecast skill. Time-series of forecast error in NINO4 are clearly correlated with ENSO phase in a way that they are not in the eastern Pacific; related to this, simple statistical processing can improve the skill of NINO4 forecasts; and there is a very large gap between RMSE and the model-estimated predictability limit. Further, past work has shown that forecast performance in the west Pacific is particularly sensitive to the model mean state (Anderson et al. 2007—see Sect. 4.6.1). It may well be that simulations from S3 are not sufficiently realistic in this region to benefit much from the improvements in the observing system. For the first month of the forecast, a noticeable reduction in NINO4 RMSE in the 1994–2007 period *is* visible, suggesting that the initialization of the model is improved, even if this does not help the longer range forecasts.

ENSO forecast skill can also be compared over longer periods of time. The S3 hindcasts were experimentally extended to cover the period 1961–80, although only with 4 start dates for each year (1st of February, May, August and November). A time-series plot of forecast error based on quarterly starts from 1961 to 2007 (not shown) indicates that errors in NINO3.4 SST forecasts were low in the 1960s and early 1970s. Plots comparing 1961–80 with 1981–2007 confirm that the earlier period had r.m.s. errors no larger than the later period, despite the very limited observing system in these early decades. However, although RMSE is not very different, both ACC and MSSS show a notable improvement of the S3 forecasts for the later decades relative to the earlier ones. The explanation is that the 1960s had only weak ENSO variability which enables forecasts to have low RMSE, but also gives poor ACC and MSSS. This is very different from our original comparison of the pre- and post- TAO eras (Fig. 6a), where ENSO variability was roughly constant, and where the later period was improved according to all scores examined.

It is also interesting to split the "post-TAO" years into two: 1994–2000 and 2001–2007 (Fig. 6b). The most recent period again has the lowest RMSE, and it might be tempting to conclude that further improvements to the observing system (e.g. improved satellite winds and ARGO floats) are responsible for the low RMSE. However, the last 7-year period has had weak ENSO variability—the RMSE of persistence is also relatively low, and S3 scores such as ACC and MSSS are worse than they were in 1994–2000. It is not feasible to disentangle the effects of the differing ENSO characteristics on the one hand, and the improvement of the observing system on the other, just by looking at the time evolution of the scores. However, observing system experiments which remove the ARGO floats from the S3 ocean analysis suggest that at least some of the recent improvement might be explained by continued improvements in the observing system (Balmaseda and Anderson 2009).

The final way in which the time evolution of skill is considered is when stratified according to season. As already noted, the prediction skill of S3 and the predictability characteristics of the equatorial Pacific are strongly seasonally dependent, and aggregating forecasts from all times of year may obscure aspects of performance. Figure 7a shows a time series of mean absolute error (MAE) for the first 3 months of the forecast, for forecasts starting

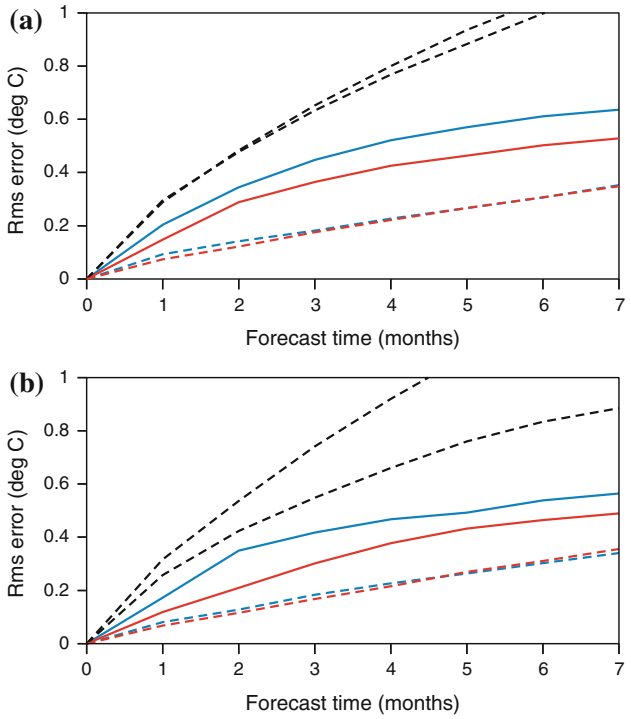

Fig. 6 RMSE for NINO3.4 forecasts from S3, for the periods **a** 1981–1993 (*blue*) and 1994–2007 (*red*), and **b** 1994–2000 (*blue*) and 2001–2007 (*red*). RMSE is lower in more recent periods, but conclusions must be carefully drawn. See text

on the 1st February, chosen because these forecasts show the greatest change in forecast skill. Figure 7b shows the RMSE of these forecasts for the years 1994-present, while Fig. 7c shows the mean bias for NINO3.4.

The time-series in Fig. 7a shows a strong reduction in error in forecasts made from 1994 onwards, with forecasts made after this date close to being "perfect", both in the average sense of the RMSE matching the ensemble spread, and in the particular sense that no forecasts have unexpectedly large errors. The change is so substantial, and occurs at such a notable change point for the observing system, that the natural hypothesis is that the forecasts improve because of the improved observing system. Even if we allow for the potential effects of selection bias (see Appendix), it is very unlikely that the dramatic improvement post-TAO can be due purely to chance.

The skill of the S3 forecasts relative to persistence is very large (Fig. 7b), but a substantial part of this is due to the decay of SST anomalies at this time of year, which the S3 forecasts capture very well. When S3 is compared with damped persistence (not shown), its advantage is smaller. The skill of damped persistence also shows some variation

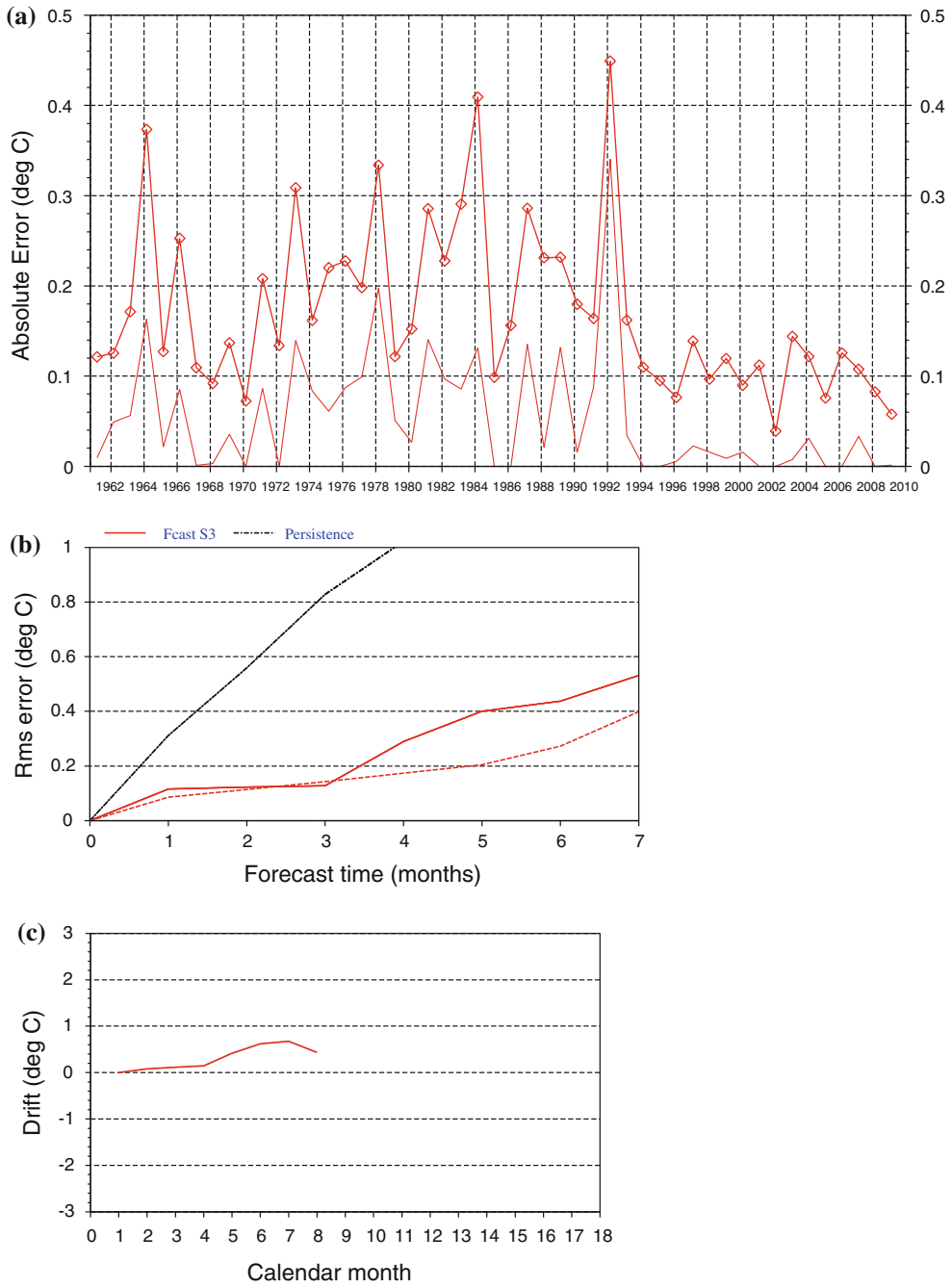

Fig. 7 **a** Time series of MAE (*thick line*) for the first 3 months of NINO3.4 forecasts starting 1st February each year. Also shown is what we call the best absolute error (BAE), which is defined at each lead time as either zero (if the observations lie within the predicted range) or the distance between the observed value and the closest ensemble member, and then averaged over lead times. For a perfect forecasting system with a modest ensemble size, the BAE would be mostly zero, with occasional small positive values. The step change in forecast skill after 1993 is evident. **b** RMS errors for NINO3.4 forecasts from 1st February 1994–2008, from S3. The RMSE for the first 3 months is small, and matches the ensemble spread well. **c** Mean bias in S3 NINO3.4 forecasts from 1st Feb, based on forecasts in the 1981–2005 calibration period. Bias is unusually small for the first 3 months

over time. From 1961 to 1992 the time evolution of MAE in the S3 forecasts and damped persistence is broadly similar, with a tendency to slightly higher errors in the 1972–1992 period, and slightly lower errors in the 1960s. From 1993 onwards, however, the collapse in errors from S3 is accompanied by only a modest decline in errors from damped persistence. The improved performance of the Feburary S3 forecasts since the start of the TAO era is much more than a fluctuation in the predictability characteristics of ENSO.

The fact that a strong and clear improvement in SST forecasts can be seen for one season from 1994 onwards, while only a weaker improvement in other seasons, deserves explanation. The impact of improved ocean initial conditions on the forecast is expected to depend on the level of error in the forecast model. Large model errors will obscure the benefit, whereas if model errors are small (for a certain time of year and a limited forecast range), then the quality of the initial conditions should be apparent. Figure 7c shows the mean bias of NINO3.4 SST for February forecasts from S3—the bias is very small for the first 3 months, being not much more than 0.1 K, although it increases at longer lead times. Figure 7b shows that the forecast RMSE matches very closely the ensemble spread for the first 3 months of the forecast. For this part of the ocean, at this time of the year, it seems that the errors from the forecast model are negligible, and the quality of the ocean initial conditions is exposed.

The seasonality of the "success" of the model appears related to the verifying period (in particular March/April) more than the start date - forecasts show particular success (and post-1994 improvement) for this verification period for a range of lead times, with the February starts being only the strongest example. Physically, March/April is the period when the NINO3.4 SST is warming most strongly in the seasonal cycle, and the physical processes controlling the evolution of SST anomalies at this time will differ from the rest of the year. It is thus plausible for a model to do well in this limited period, while still having significant errors at other times. However, success in March/April forecasts is by no means assured. For example, S2 had significant biases at this time of year and showed neither a strong time-dependence nor near-perfect forecast skill in recent years.

4.5 Skill of 13 month forecasts

Figure 8 shows the RMSE and ACC of NINO3.4 forecasts out to 13 months, for the years 1981–2007. The ensemble spread is also plotted, scaled by a factor of $\sqrt{(6/5)}$ to give the predictability limit of the forecast system for a 5 member ensemble, to allow proper comparison between actual and potential performance. In the operational system

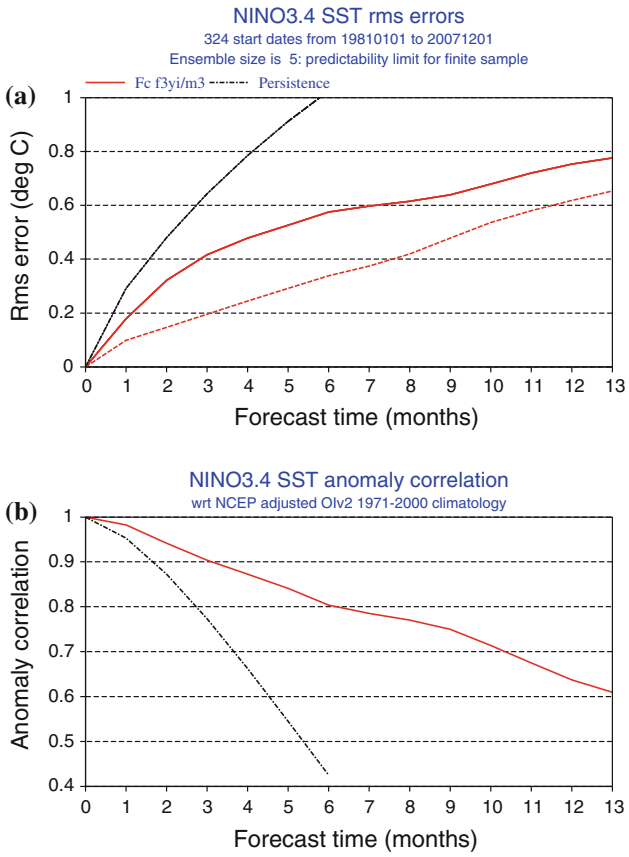

Fig. 8 RMSE and ACC for NINO3.4 SST for annual range forecasts, initialized every month for the years 1981–2007. Skill extends to a lead time of 1 year, and at longer leads the ensemble spread (*dashed line*) begins to approach the actual RMS error

these forecasts are only made every 3 months, but additional experimentation has since been performed to fill in the missing months and give a more complete picture of skill. Two features stand out: the ACC skill remains high over the full period, above 0.75 for the first 9 months and still above 0.6 at 13 months. Also, RMSE grows quickly in the early months and then slows, with a slight kink upwards after 9 months, while the ensemble spread grows steadily throughout, leading to a reduced gap between actual and potential performance at longer lead times: model error becomes relatively less important. When plotted in terms of error variance (not shown), the ensemble variance is tiny at first, but explains a rapidly increasing fraction of the total error variance and dominates the error budget from 9 months onwards. This may explain the "kink" at 9 months: model-related errors grow rapidly in the first few months but then hardly at all, while errors associated with the predictability limit grow continuously, being negligible at first but eventually becoming large.

These results demonstrate that (i) there is significant forecast skill for ENSO related SST at timescales of a year or more ahead; (ii) S3 does a good job of predicting SSTs

on these timescales; and (iii) inherent noise and error growth in the system means that longer range ENSO SST forecasts have an increasing and unavoidable uncertainty, and the need to consider them in a probabilistic rather than deterministic way becomes unavoidable.

Predictability of ENSO by a coupled GCM at these timescales has also been reported by Luo et al. (2005, 2008), although their use of a 5 month smoothing before verification precludes direct comparison of forecast skill between the two systems. It should also be noted that at longer leads, where predictability is lower, the scores are moderately sensitive to sampling of forecasts, even in a common 27 year period. For instance, if we take just the S3 operational forecasts (at 3 month intervals) instead of the full set, the ACC at 13 months exceeds 0.7. Also, the fact that only a 5 member ensemble is available, impacts negatively on S3 skill at longer leads when the ensemble spread is large. A bigger ensemble size would give systematically better scores. Implementing a larger ensemble operationally is simply a resource issue.

5 Forecast skill in the Indian and Atlantic Oceans

The skill of S3 forecasts of SST in the tropical Indian and Atlantic oceans is now considered. Correlations between ensemble-mean and observed SST anomalies, averaged over specific regions of the two oceans, have been computed from the 1981–2005 hindcasts, and are presented in Fig. 9 as a function of verification month (*x*-axis) and forecast lead time (*y*-axis).

The top panels in Fig. 9 refer to the two regions used to define the so-called Indian Ocean Dipole (IOD; e.g. Saji et al., 1999; Webster et al. 1999); these are usually referred to as western tropical Indian Ocean (WTIO; top-left) and south-eastern tropical Indian Ocean (SETIO; top-right). For both regions, SST anomaly correlation remains high (>0.6) even at long lead times for forecasts verifying in the first half of the year. In the case of the SETIO, this is in stark contrast to persistence, which has no skill in predictions for the first half of the year which are initialized before January. The model suffers a significant drop in skill for predictions

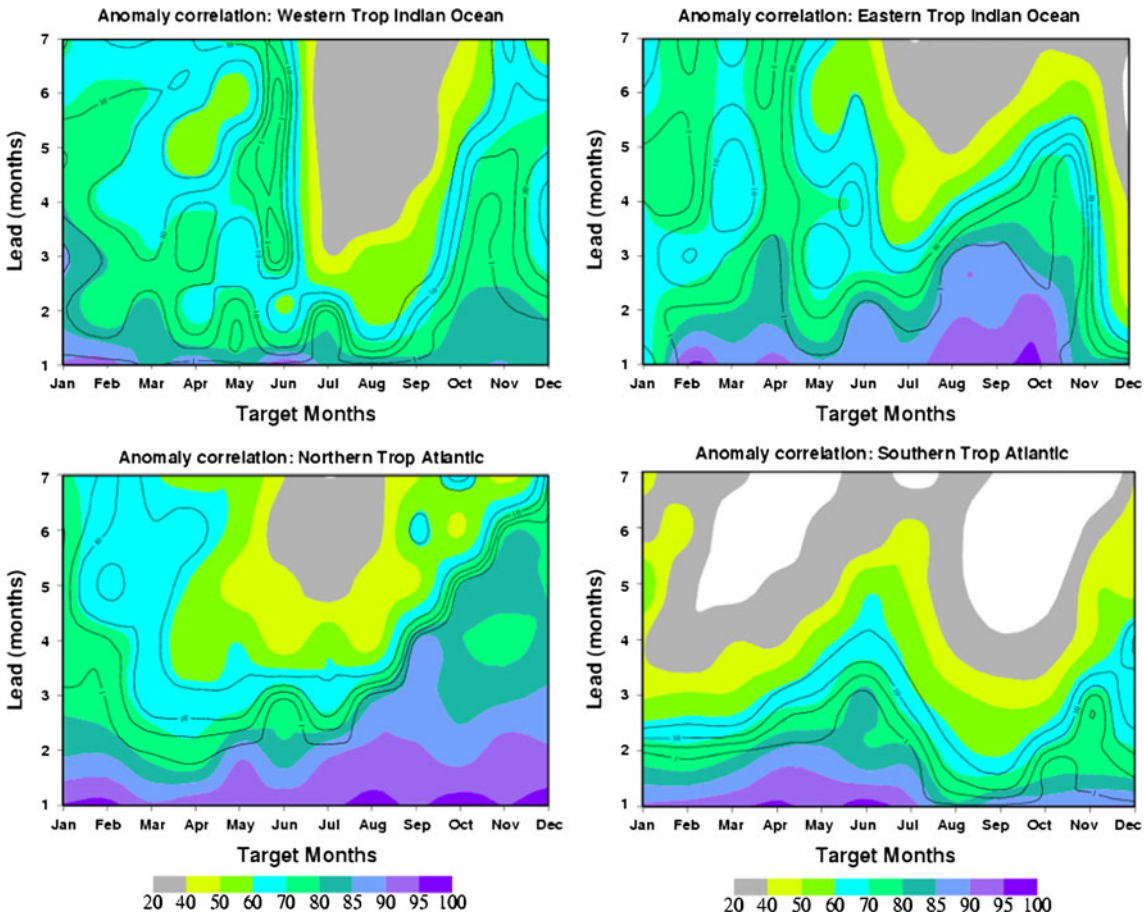

Fig. 9 Correlation between ensemble-mean anomalies of SST and observed anomalies as a function of verification time (*x*-axis) and forecast lead time (*y*-axis) in months, for four different regions in the Indian and Atlantic oceans. *Top left*: western tropical Indian Ocean (50E–70E, 10S–10N); top right: south-eastern tropical Indian ocean. (10S-0, 90E–110E); *bottom left*: northern tropical Atlantic (60W–15 W, 5N–25N); *bottom right*: southern tropical Atlantic (20S-0, 30 – 15E)

verifying in early summer at long lead times, especially for the western area (seasons refer to NH). Short lead time predictions for SETIO show a very high correlation skill in the August–October period, when this region is usually anti-correlated with WTIO, and the Indian Ocean Dipole reaches its largest amplitude. In WTIO during autumn, forecast skill at longer leads recovers from its summer minimum, and this leads to skilful predictions of the IOD in this season (not shown). In late winter and spring, on the other hand, variability in the tropical Indian Ocean tends to be positively correlated across the basin (e.g. as a response to ENSO). This makes the IOD index less relevant, and its predictions are much less skilful than in the second half of the year, despite the fact that the individual WTIO and SETIO regions have relatively high predictability.

The lower panels in Fig. 9 refer to regions representing the North and South Tropical Atlantic. In both regions the S3 seasonal forecasts are a moderate improvement on persistence (not shown), more obviously so in the northern part of the dipole (lower left panel in Fig. 9). The S3 forecasts are at their best when initialized in June or later and predict values through to the end of the year. The skill of S3 forecasts verifying in spring is also moderately high (and higher than persistence) this being the season when the influence of ENSO is stronger in this area, which is well captured by the model (not shown). The lowest level of skill, only marginally better than persistence, is for predictions verifying during the NH summer.

Forecasts over the South Tropical Atlantic (lower right panel in Fig. 9) generally lose skill rapidly after the first 2 or 3 months. The highest values are for predictions initialized in SH late summer (Feb–March) and Autumn, verifying up to July. Predictions verifying during Jun-July are presumably benefiting from the correct capture by the model of an ENSO teleconnection to South Tropical Atlantic SST which is observed at this time. Why the model performance drops so rapidly for verifications in August and beyond is not clear—performance is worse than anomaly persistence at this time.

The prediction of equatorial SST anomalies in the Atlantic (not shown) is slightly improved compared with previous results (Stockdale et al. 2006), but the overall picture is similar. In particular, despite some skill in predicting boreal winter SST, prediction for JJA is poor except at short leads, and offers no real benefit over persistence. Model biases remain a challenge in this region, particularly in the boreal summer, and it is unsurprising that forecast skill is still limited.

6 Concluding discussion

The latest seasonal forecast system at ECMWF, System 3, has been described, and its ability to predict SST has been catalogued. The forecast skill is higher than any previous ECMWF operational system, and in the tropical Pacific the SST is generally well predicted. In particular, the short range forecasts (0–1 months lead) are of high quality, and beat persistence of anomalies by a large margin. Skill relative to both persistence and climatology remains good at longer leads, even out to 12 months. Onset, intensification and peak of El Nino and La Nina events are generally well predicted in the 6 month lead forecasts, although there are occasions when the model develops a warm event too readily (e.g. 1990). Forecast skill for Nino 3.4 SST is high throughout the year, and thus shows big gains over persistence for verification in the May–October period. Skill remains below the estimated predictability limit, suggesting there is still scope to improve forecast scores in the future.

One interesting result concerns the variation of ENSO forecast skill over time. The skill in more recent periods, particularly post 1994, is appreciably higher than in earlier periods. As discussed in the text, there are a number of complications when comparing ENSO skill across different time periods—the impact of ENSO activity on skill measures such as RMSE and ACC, and the possibility of regime-dependent model errors affecting forecasts in specific periods. Nonetheless, careful analysis suggests that there is a real improvement in skill in the post-TAO years relative to the preceding 13 year period, which cannot be explained trivially by changes in ENSO characteristics.

A straightforward explanation of these results is that the improved observing system, notably the TAO array itself and also altimetry, is responsible for a large part of the forecast improvement. Indeed, oceanographic observing system experiments (Balmaseda and Anderson 2009) have established that the "new" observing systems of TAO ocean data and altimeter do improve the S3 forecasts.

A full assessment of the impact of the complete observing system on forecast skill is possible, but not straightforward. In particular, an important part of the TAO measurements are of surface winds, which feed into the NWP (re-)analyses used to drive the ocean assimilation system. Indeed, earlier unpublished work with a multi-model ensemble from the DEMETER project (Palmer et al. 2004) shows noticeable improvements in short-range NINO3.4 forecast skill post 1994, despite the models concerned being initialized only with winds and SST, and not sub-surface temperature data. This result suggested that improved specification of the winds in the TAO era gives better forecasts, if a multi-model approach is used to reduce the impact of model error on the forecasts. A rigorous assessment of the impact of the full observing system would require observing system experiments involving atmospheric as well as oceanic reanalyses, which would be a major undertaking.

The most striking time-variation of forecast skill is obtained by looking only at forecasts starting in February,

and verifying them only for the first 3 months. The result is so strong that it is not an artifact of data selection, but rather appears to be a consequence of model errors being small at a particular time of year, thus allowing better observations to have a clearer impact on ENSO forecast skill. This fits alongside the assessment that the level of error in the forecasts is typically much larger than expected from estimates of the uncertainty in the initial conditions and the predictability of the system, and that model error is therefore usually dominant.

The view that model error, contaminating both analyses and forecasts, dominates today's forecast errors, has several consequences. It explains why the sensitivity of forecast skill to the improvement of observations is often weak, and suggests that there is still much scope for improving forecasts with better models. Further, the value of existing observing systems for improving forecasts is likely to increase as models are improved. The results from the S3 February forecasts offer a tantalizing glimpse of what may one day be possible.

Outside of the equatorial Pacific, S3 has both successes and failures. SST in the equatorial Indian Ocean is generally well predicted, with particular success in the eastern part of the basin, and in predicting the Indian Ocean Dipole index in boreal autumn. The equatorial and southern tropical Atlantic, however, remain difficult regions, and there are clearly issues with high latitude winter SST in the southern hemisphere.

Overall, it can be concluded that the development and implementation of the ECMWF S3 seasonal forecasting system has been a substantial and significant achievement, setting new standards for the prediction of ENSO related SST. It will be a challenge for the future to develop a "System 4" forecast system which can outperform S3. Despite the successes so far, there are still many improvements that are needed. In some cases, the way forward is clear: it is intended to make improvements to the initialization of soil moisture, which is important for 2-m temperature in the early seasonal range; and as resources allow, it is intended to implement better representations of stratospheric forcings and sea-ice. More generally, though, the ability of a coupled model to accurately represent all aspects of ENSO-related SST variability, and to capture the correct atmospheric responses, remains as a real challenge to climate modellers on all timescales.

Acknowledgments The improvements in the atmospheric model used in S3 are due to dedicated work by many individuals at ECMWF.

Appendix

If a specific season and lead time have been selected because they give the best results, it is necessary to ensure that what is seen is not just a sampling effect. As a simple combinatorial test, define a "poor" forecast as one in which the mean absolute error for the first 3 months exceeds 0.15° C. If we consider all the S3 forecasts from 1961 to 2009 present (49 years), a total of 23 of them are poor, all in the years prior to 1994. If it is assumed, as a null hypothesis, that the poor forecasts are equally likely to occur in any year, the chances of the last 16 years including zero occurrences of the 23 poor forecasts which exist in the sample is (33/49)*(32/48)* …(11/27) or (33!/49!)*(26!/10!) = 0.0000016. Using this method of looking for a change in skill (defining thresholds, counting poor forecasts, and applying combinatorial tests to get a p value), the following selections have been made: the best of 12 start months, the best of 7 possible forecast ranges, and the best of (say) 5 plausible thresholds for the definition of a poor forecast. This multiplies to 420 possible tests for a change in skill, of which this is the highest scoring. Under the null hypothesis, the chances of such a high score are only 0.0007.

References

Anderson D, Stockdale T, Balmaseda M, Ferranti L, Vitart F, Molteni F, Doblas-Reyes F, Mogenson K, Vidard A (2007) Development of the ECMWF seasonal forecast System 3. ECMWF Technical Memoranda 503

Balmaseda M, Anderson D (2009) Impact of initialization strategies and observations on seasonal forecast skill. Geophys Res Lett 36:L01701. doi:10.1029/2008GL035561

Balmaseda MA, Dee D, Vidard A, Anderson DLT (2005) A multivariate treatment of bias for sequential data assimilation: application to the tropical oceans. Q J Roy Meteorol Soc 133:167–179

Balmaseda MA, Vidard A, Anderson D (2008) The ECMWF ORA-S3 ocean analysis system. Mon Wea Rev 136:3018–3034

Balmaseda MA, Ferranti L, Molteni F, Palmer TN (2010) Impact of 2007 and 2008 Arctic ice anomalies on the atmospheric circulation: implications for long-range predictions. Q J Roy Meteor Soc 136:1655–1664. doi:10.1002/qj.661

Buizza R, Miller M, Palmer TN (1999) Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. Q J Roy Meteor Soc 125:1908–2887

Cane MA, Zebiak SE, Dolan SC (1986) Experimental forecasts of El Nino. Nature 321:827–832

Doblas-Reyes FJ, Hagedorn R, Palmer TN, Morcrette J-J (2006) Impact of increasing greenhouse gas concentrations in seasonal ensemble forecasts. Geophys Res Lett 33:L07708. doi:10.1029/2005GL025061

Hurrell J, Meehl GA, Bader D, Delworth TL, Kirtman B, Wielicki B (2009) A unified modeling approach to climate system prediction. Bull Am Meteorol Soc 90:1819–1832

Jin EK, Kinter JL III, Wang B, Park C-K, Kang I-S, Kirtman BP, Kug J-S, Kumar A, Luo J-J, Schemm J, Shukla J, Yamagata T (2008) Current status of ENSO prediction skill in coupled ocean-atmosphere models. Clim Dyn 31:647–664. doi:10.1007/s00382-008-0397-3

Luo J-J, Behera S, Shingu S, Yamagata T (2005) Seasonal climate predictability in a coupled OAGCM using a different approach for ensemble forecasts. J Clim 18:4474–4497

Luo J-J, Masson S, Behera SK, Yamagata T (2008) Extended ENSO predictions using a fully coupled ocean-atmosphere model. J Clim 21:84–93

McPhaden MJ (2003) Tropical Pacific Ocean heat content variations and ENSO persistence barriers. Geophys Res Lett 30:1480. doi: 10.1029/2003GL016872

Meehl GA, Arblaster JM, Branstator GW, van Loon H (2008) A coupled air-sea response mechanism to solar forcing in the Pacific region. J Clim 21:2883–2897

Palmer TN, Alessandri A, Andersen U, Cantelaube P, Davey M, Délécluse P, Déqué M, Díez E, Doblas-Reyes FJ, Feddersen H, Graham R, Gualdi S, Guérémy J-F, Hagedorn R, Hoshen M, Keenlyside N, Latif M, Lazar A, Maisonnave E, Marletto V, Morse AP, Orfila B, Rogel P, Terres J-M, Thomson MC (2004) Development of a European multi-model ensemble system for seasonal to inter-annual prediction (DEMETER). Bull Am Meteorol Soc 85:853–872

Palmer TN, Doblas-Reyes FJ, Weisheimer A, Rodwell MJ (2008) Toward seamless prediction: calibration of climate change projections using seasonal forecasts. Bull Am Meteor Soc 89:459–470

Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP, Kent EC, Kaplan A (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J Geophys Res 108:4407. doi: 10.1029/2002JD002670

Reynolds RW, Rayner NA, Smith TM, Stokes DC, Wang W (2002) An improved in situ and satellite SST analysis for climate. J Clim 15:1609–1625

Saha S, Nadiga S, Thiaw C, Wang J, Wang W, Zhang Q, Van den Dool HM, Pan HL, Moorthi S, Behringer D, Stokes D, Peña M, Lord S, White G, Ebisuzaki W, Peng P, Xie P (2006) The NCEP climate forecast system. J Clim 19:3483–3517

Saji NH, Goswami BN, Vinayachandran PN, Yamagata T (1999) A dipole mode in the tropical Indian Ocean. Nature 401:360–363

Stockdale TN (1997) Coupled ocean atmosphere forecasts in the presence of climate drift. Mon Wea Rev 125:809–818

Stockdale TN, Anderson DLT, Alves JO, Balmaseda MA (1998) Global seasonal rainfall forecasts with a coupled ocean atmosphere model. Nature 392:370–373

Stockdale TN, Balmaseda MA, Vidard A (2006) Tropical Atlantic SST prediction with coupled ocean-atmosphere GCMs. J Clim 19:6047–6061

Tompkins AM, Feudale L (2010) Seasonal ensemble predictions of West African monsoon precipitation in the ECMWF System 3 with a focus on the AMMA special observing period in 2006. Weather Forecast 25:768–788

van Oldenburgh GJ, Balmaseda M, Ferranti L, Stockdale T, Anderson D (2005) Did the ECMWF seasonal forecast model outperform statistical ENSO forecast models over the last 15 years? J Clim 18:3240–3249

Vialard J, Vitart F, Balmaseda M, Stockdale T, Anderson D (2005) An ensemble generation method for seasonal forecasting with an ocean-atmosphere coupled model. Mon Weather Rev 133: 441–453

Wang B, Lee J-Y, Kang I-S, Shukla J, Park C-K, Kumar A. Schemm J, Cocke S, Kug J.-S, Luo J-J, Zhou T, Wang B, Fu X, Yun W-T, Alves O, Jin EK, Kinter J. Kirtman B, Krishnamurti T, Lau NC, Lau W, Liu P, Pegion P, Rosati T, Schubert S, Stern W, Suarez M, Yamagata T (2009) Advance and prospectus of seasonal prediction: assessment of the APCC/CliPAS 14-model ensemble retrospective seasonal prediction (1980–2004). Clim Dyn. doi: 10.1007/s00382-008-0460-0

WCRP (2005) The world climate research programme strategic framework 2005–2015: coordinated observation and prediction of the earth system (COPES). WCRP-123, WMO/TD-No. 1291, 65

Webster PJ, Moore A, Loschnigg J, Leban M (1999) Coupled ocean-atmosphere dynamics in the Indian Ocean during 1997–98. Nature 401:356–360

Weisheimer A (2005) SST and wind stess perturbations for seasonal and annual simulations. Available from: http://www.ecmwf.int/research/EU_projects/ENSEMBLES/exp_setup/ini_perturb/index.html

Weisheimer A, Doblas-Reyes FJ, Palmer TN, Alessandri A, Arribas A, Déqué M, Keenlyside N, MacVean M, Navarra A, Rogel P (2009) ENSEMBLES: a new multi-model ensemble for seasonal-to-annual predictions—skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. Geophys Res Lett 36:L21711. doi:10.1029/2009GL040896

Zwiers FW, von Storch H (1995) Taking serial correlation into account in tests of the mean. J Clim 8:336–351