TECHNICAL MEMORANDUM

835

SEAS5 and the future evolution of the long-range forecast system

Tim Stockdale, Magdalena Balmaseda, Stephanie Johnson, Laura Ferranti, Franco Molteni, Linus Magnusson, Steffen Tietsche, Frederic Vitart, Damien Decremer, Antje Weisheimer, Christopher Roberts, Gianpaolo Balsamo, Sarah Keeley, Kristian Mogensen, Hao Zuo, Michael Mayer and Beatriz Monge-Sanz

Research Department, Forecast Department and Copernicus Department

Presented to the Scientific Advisory Committee, October 2018

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/en/research/publications

Contact: library@ecmwf.int

European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, Berkshire RG2 9AX, England

The information within this publication is given in good faith and considered to be true, but ECMWF accepts no liability for error, omission and for loss or damage arising from its use

Contents

Abstract

SEAS5 became operational in November 2017, following almost two years of effort dedicated to the upgrade. To date, the seasonal forecasting configuration at ECMWF has been upgraded only infrequently, so each new version represents a major change, and SEAS5 is no exception. As well as many scientific developments, SEAS5 highlights include major increases in resolution for both atmospheric and oceanic models, the introduction of an active sea-ice model, and a shift in strategy to align the seasonal forecast configuration more closely to that used for medium-range forecasts. ENSO SST forecasts, which were already good, show further substantial improvement. Many other aspects of forecast performance are also improved, although score differences are often within sampling error and a few specific areas of deterioration have also been identified.

This paper gives some background to the development of SEAS5 (Section 1) and a description of the SEAS5 configuration (Section 2). Section 3 gives an assessment of performance, including various studies to help understand the factors affecting forecast performance. Finally, Section 4 is forward looking, with an overview of scientific priorities and a roadmap for future developments, including SEAS6.

1 Introduction

Work on seasonal predictability started at ECMWF in the 1980s, and in 1993 ECMWF Council approved an experimental programme in seasonal prediction. This led to the development of a coupled ocean-atmosphere seasonal forecasting system which started real-time running in 1997, just in time to successfully predict the 1997 El Niño and many of its impacts. This first seasonal forecast system ran until it was replaced in January 2002 by “System 2”, and the seasonal forecasting system has been upgraded at approximately 5-year intervals ever since, with System 3 in March 2007 and System 4 (or S4) in November 2011. SEAS5 replaced S4 in November 2017, twenty years after the first ECMWF real-time seasonal forecast. The change in nomenclature reflects wider changes at ECMWF: we have a single integrated forecast system (IFS) in different configurations, HRES, ENS and SEAS.

During the last 20 years there have been considerable advances in seasonal forecasting systems. The horizontal resolution in SEAS5 is about 8 times that in the system 1; the vertical resolution in the upper ocean is now about 20 times finer than twenty years ago. Continuous coupled model development means that key processes like the Madden Julian Oscillation, a major challenge two decades ago, are now represented by the coupled model. There has also been tremendous progress in atmospheric and ocean reanalyses, which allow for longer reforecast records with more accurate initial conditions. Major design and infrastructure developments have also contributed to the consolidation of the seasonal forecasting systems. Section A.1 in the Appendix provides more details on the evolution of seasonal forecasting at ECMWF. Scientific understanding and user uptake has benefited from the involvement in EU projects such as DEMETER and ENSEMBLES. At ECMWF, the legacy of 20 years of seasonal activities is clearly visible: i) the maturity of seasonal forecasting manifests in the provision of multi-model seasonal forecasting by the Copernicus Climate Change Services (C3S); ii) the coupled model and ocean re-analysis, initially a prerogative of the seasonal range, are now fully integrated in seamless forecasting systems.

The seamless approach implies that the practices adopted for development and implementation of the seasonal system need revisiting. The seasonal system has been upgraded much less frequently than the medium-range forecast system, for several reasons. In the early years of seasonal prediction the forecasting system was of necessity very different from that used in the medium-range. Seasonal forecasting required an ocean model and ocean analysis and a suitable representation of a wider range of processes - although the seasonal system was always based on a specific cycle of the IFS, it was in practice a very separate system. Beyond this, it became clear after S2 that new IFS cycles could be problematic, and that by upgrading the seasonal system only occasionally there was more control to implement only "suitable" cycles of the IFS. There was also a belief that the developing user community preferred stability, strengthened by requests to keep old systems running for some time after new ones were introduced. Finally, seasonal forecasting required, and still requires, a substantial set of re-forecasts, which are expensive and time-consuming to produce. Lengthy gaps between new systems were necessary to allow scientific and technical development with the limited resources available. The extent to which these considerations are still valid is discussed in Section 4.

Section 2 of this paper provides a summary description of SEAS5; a more detailed description is available in Johnson et al (2018). Section 3 gives a detailed analysis of the performance of SEAS5, both in terms of forecast scores and physical processes. Finally, Section 4 looks to the future, considering both specific scientific issues and our seamless strategy.

2 Description of SEAS5

Table 1 lists the main features of the SEAS5 model and its predecessor, S4. Notable upgrades in SEAS5 w.r.t. S4 are the substantial increase in atmospheric and ocean resolution, and the inclusion of a prognostic sea-ice model. SEAS5 is now much more seamless with ENS than ever before. The reforecast length and ensemble size has also increased. The different aspects of the SEAS5 forecast system - forecast model, initialization, ensemble generation, reforecast and real-time production – are described below.

2.1 Model Configuration

IFS horizontal and vertical resolution is now the same as used in the extended-range ENS configuration. This is deliberate and helps minimize the number of different configurations in operational use. We detail here the small number of changes made to SEAS5 to ensure the model was appropriate for longer range forecasts. In other aspects SEAS5 is identical to ENS.

In SEAS5 the tropical amplitude of the parametrized non-orographic gravity wave drag was considerably reduced compared to the default settings in 43r1, the reduction being necessary to give reasonable behaviour of the Quasi-Biennial Oscillation. The re-tuned parameters were included in the medium-range ENS system from Cy45r1. This is a nice example of the seamless strategy not being a strict constraint on individual model cycles, but a process that can drive model improvement for all timescales.

SEAS5 has a few small modifications to radiative forcing, to allow representation of long term variability relevant for proper calibration of the real-time forecast relative to previous decades. Tropospheric sulphate aerosol follows ERA5, using the decadally varying CMIP5 climatology rather than the fixed default 43r1 climatology. This is less important but still relevant for medium-range forecasts and, in the future, it should be possible to find a unified treatment of this effect for all our forecast configurations. Note that black carbon in SEAS5 is still the fixed 43r1 climatology, and tropospheric aerosols are still considered an area in need of further development. SEAS5 retains the S4 treatment of volcanic stratospheric sulphate aerosol, with damped persistence of an initially specified loading. Details are given in Appendix A.2.

In line with the seamless strategy, prognostic ozone is not radiatively interactive as it was in S4. Instead, the radiation scheme sees the same ozone climatology used in the 43r1 ENS extended forecasts. In future cycles, it is envisaged to reintroduce the interaction of prognostic stratospheric ozone with radiation alongside developments for improving the representation of the stratosphere.

Table 1 A comparison of the forecasting system configurations in S4 and SEAS5

	S4	**SEAS5**
	Coupled Model	
IFS Cycle	36r4	43r1
IFS resolution (TOA)	TL255 (80 Km) L91 (0.01 hPa)	TCo319 (36 Km) L91 (0.01 hPa)
Ocean model and	NEMO v3.0	NEMO v3.4
Ocean model resolution	ORCA 1.0 Z42 (10m upper level)	ORCA 0.25 Z75 (1m upper level)
Sea ice model	Sampled climatology	LIM2
Wave model resolution	1.0	0.5
Coupler	OASIS	Single Executable
	Ensemble Generation	
Ocean Initial Perturbations	ORAS4 5-members + SST perturbations	ORAS5 5-members + SST perturbations. Updated Scheme
Atmosphere Initial Perturbations	SV	SV + EDA
Model stochastic physics	3-scale SPPT and SKEB	3-scale SPPT (conservation fix) and SKEB
	Reforecasts and initial conditions	
Period and Ensemble members	1981-2010. 15 members	1981-2016. 25 members
Ocean and Sea Ice initial conditions	ORAS4 [SST: OIv2+OSTIA; in-situ: EN4]	ORAS5 [SST:HadISSTv2+ OSTIA; in-situ: EN4]
Atmosphere initial conditions	ERA-I	ERA-I
Land Initial conditions	HTESSEL TL255 Cy36r4 forced by ERAI (GPCP corrected). No lakes	HTESSEL TCo319 Cy43r1 forced by ERAI (uncorrected). It includes lakes.
	Real Time forecasts and initial conditions	
Ensemble members	51	51
Initialization	As ENS	As ENS
Release date	8th of each month	5th of each month

2.2 Ocean, sea-ice and wave model

SEAS5 uses the NEMO ocean model, upgraded to version 3.4, with horizontal resolution increased from ORCA1 (1°) in S4 to ORCA025 (0.25°), and the number of vertical levels increased from 42 to 75. Near surface resolution is particularly increased: the depth of the top layer decreases from 10 metres to 1 metre. This new NEMO configuration is the result of a pan-European collaborative effort, involving the

DRAKKAR consortium and the MetOffice and NOCS (National Oceanographic Centre in Southampton).

SEAS5 now has an active sea-ice model, using LIM2 which is part of the NEMO modelling framework. The implementation of the LIM2 model and its coupling with the IFS has benefited from collaborations with EC-Earth. The sea-ice model is fully coupled and exchanges mass, heat, and momentum with the atmosphere above and the ocean below. The ice is single-category but includes two layers of ice and a snow layer. The ice dynamics uses the Hibler visco-plastic rheology. The biggest limitation of the sea-ice coupling with the IFS is that the surface fluxes calculated by the IFS (and subsequently used by the ice model to drive ice growth rates) are calculated using a constant 1.5m ice thickness in the IFS. This means that ice growth rates are too slow when ice is thin, and too fast when ice is thick. Nonetheless, this prognostic sea-ice model allows sea-ice cover to respond to changes in the atmosphere and ocean states, enabling SEAS5 to provide seasonal outlooks of sea-ice cover.

The wave model also benefits from a resolution increase, and the physics of the atmosphere-ocean-wave coupling are more comprehensive than previously. The coupling between IFS and ocean no longer uses the OASIS3 coupler, but relies on a coupling interface within the IFS, with the whole model being run as a single executable.

2.3 Initialization

SEAS5 ocean and sea-ice initial conditions for forecasts and reforecasts are provided by a new operational ocean analysis system (OCEAN5) made up of the historical ocean reanalysis (ORAS5) and the daily real-time ocean analysis (OCEAN5-RT). OCEAN5 uses the same ocean and sea-ice model as the coupled forecasts in SEAS5. Compared to its predecessor ORAS4 (Balmaseda et al., 2013), OCEAN5 has higher resolution, updated data assimilation and observational data sets and provides sea-ice initial conditions.

ORAS5 is based on Ocean Reanalysis Pilot 5 (Tietsche et al, 2017; Zuo et al, 2017b, ORAP5), but uses updated observational data sets. The ocean in-situ temperature and salinity come from the recent quality-controlled EN4 (Good et al, 2013), which has higher vertical resolution and fuller spatial coverage than the previous version EN3. The altimeter sea-level data have also been updated to the latest version (DUACS2014) from CMEMS (Copernicus Marine Environmental Monitoring Services). The underlying SST analysis before 2008 now comes from the HadISST2 dataset, the same used in the ERA5 atmospheric reanalysis. The sea-ice concentration before 1985 comes from ERA-40 and from 1985 to 2008 it comes from an OSTIA reprocessed product. From 2008 onwards, the SST and sea-ice are given by the OSTIA product delivered in real-time, which is also used in the ECMWF operational analysis. More details on the system configurations and sensitivities are given in Zuo et al (2018). The fundamental upgrades of the ocean data assimilation system that have enabled ORAS5 are a product of a tight collaboration with the NEMOVAR consortium. The development of ORAS5 and its predecessor ORAP5 have benefited from the EU-funded project My-Ocean2, collaboration with the international CLIVAR and GODAE-OceanView communities, and engagement with CMEMS. Computer resources were provided by C3S.

Atmosphere initial conditions come from ERA-Interim for the re-forecasts and from ECWMF operational analyses for forecasts, from 1 Jan 2017 onwards. For the land surface, including lakes, the

HTESSEL land surface model is run offline to create initial conditions for the re-forecast period, with the same model version and resolution as SEAS5. ERA-Interim forcing is used, without any GPCP-based correction to precipitation. For the forecasts, the land surface is initialised from ECMWF operational analysis. This involves interpolation from O1280 to the O320 grid, and the analysis can also be incompatible with the offline model run. Consequently, a limiter is used to prevent the real-time land surface values taking inconsistent values relative to those used in the reforecasts. The limits are defined as the maximum and minimum values observed at that point and calendar date for the 36-year reforecast period, plus an additional margin specified as a global constant for each field. More details are given in the SEAS5 user guide.

2.4 Ensemble Generation

Atmosphere initial conditions are perturbed the same way as in ENS. Upper air fields and a limited set of land fields (soil moisture, soil temperature, snow, sea-ice temperature and skin temperature) are perturbed, using perturbations from singular vector computations plus an ensemble of data assimilations (EDA). EDA perturbations are not available for earlier years in the reforecast set, so to preserve consistency across the re-forecast set, the EDA perturbations from 2015 were applied to all re-forecast years, repeating the annual cycle each year. Use of the EDA gives SEAS5 much bigger initial perturbations in the tropics than S4.

To sample ocean state uncertainty, ORAS5 contains a 5-member ensemble analysis. This makes use of perturbations to the assimilated observations, both at the surface and at depth, and perturbations to the surface forcing fields (Zuo et al, 2017a). A larger set of SST perturbations are then applied prior to the start of each ensemble forecast. Perturbations are drawn from the ORAS5 HadISST2 pentad analysis error repository (Zuo et al, 2017a, Section 4), and applied to the upper 22 levels of the sea temperature, decreasing with depth. This is the same technique as S4, but with a different SST uncertainty dataset: the amplitude of the initial SST perturbations is smaller in the east-central equatorial Pacific, larger in the far west Pacific, broadly similar in mid-latitudes.

The atmospheric forecast model is perturbed stochastically, using identical schemes to the medium-range *(*Leutbecher et al, 2017; IFS, 2016), namely the Stochastically Perturbed Physical Tendency (SPPT) and Stochastically Kinetic Energy Backscatter (SKEB) schemes, consistent with 43r1. SPPT includes a mass, energy and moisture conservation fix that was originally developed by the EC-Earth consortium (Davini et al 2017). No stochastic physics perturbations are applied in the ocean or the land surface.

2.5 Reforecast and real-time forecast production

The operational configuration of the forecast system is also important. Some aspects are carried over unchanged from S4, but there are also some key changes:

- Release date is brought forward from the 8th to the 5th of the month, at 12Z
- Re-forecast ensemble size is increased from 15 to 25
- Re-forecast period has been extended to 1981-2016 (36 years), compared to 1981-2010 (30 years)

Although the verification uses the full 1981-2016 re-forecast period, operational charts of fields such as 2m-temperature and precipitation are presented as anomalies relative to the more recent 1993-2016 period. We consider this a better way to present seasonal anomalies to our users, since it better relates to the recent past for fields such as temperature. The 24-year reference period still offers good stability for mean climate, and is consistent with how Copernicus Climate Change Service (C3S) presents its new multi-model forecasts. For users who want to calibrate and reference the SEAS5 forecasts in ways specific to their own application, the full 36 years of re-forecast data remains available. Another change to the operational charts is the addition of SST anomaly plumes for the NINO1+2 region, important for Peru and Ecuador.

The high resolution of SEAS5 and its extensive set of re-forecasts were made possible by a significant contribution towards the costs from C3S, due to SEAS5 being one of the core contributions to the new C3S multi-model seasonal forecasting service. There has also been a change in data policy: although ECMWF retains ownership and control of the full-resolution real-time forecasts, both the re-forecast dataset and a comprehensive 1-degree resolution dataset from the real-time forecasts is publicly distributed by C3S, with a release data of the 10th of each month. Open access will lead to increased use of our raw model forecast data, and enhanced feedback from the global scientific and user communities.

3 SEAS5 Performance

3.1 Forecast Performance

3.1.1 Forecast of ENSO and other Tropical SST indices.

ENSO is the biggest source of predictability on seasonal timescales, and the quality of ENSO forecasts underpins our seasonal forecasting system. ENSO forecast quality depends on both the model and the initial conditions. Our previous systems have been world leading in terms of ENSO forecast (e.g. Barnston et al, 2012), and we would like to maintain this, while recognizing that our baseline skill is already very high.

Figure 1 shows the root mean squared error (RMSE), anomaly correlation, amplitude of variability, and seasonal variation of correlation of NINO3.4 SSTs in the central tropical Pacific from SEAS5 and S4, over 432 cases from 1981 to 2016. Sampling uncertainty in the statistics is small, due to the large number of cases considered; panel (a) includes the 95% confidence interval on the RMSE curve for SEAS5, derived from a sampling method applied to the ensemble members, but it is barely visible unless the figure is enlarged. The statistics show that SEAS5 gives consistent improvements in all these measures at almost all times of year. We note that the substantial reduction in RMSE is matched by a reduction in the ensemble spread (dashed lines in panel a), such that the forecast error remains clearly larger than the ensemble spread. At the longest leads, this problem is in fact slightly worse than it was in S4. Note that the improvement in correlation for individual target months from August through to March (Figure 1, panel d) is particularly pleasing, since this is independent of the amplitude of variability.

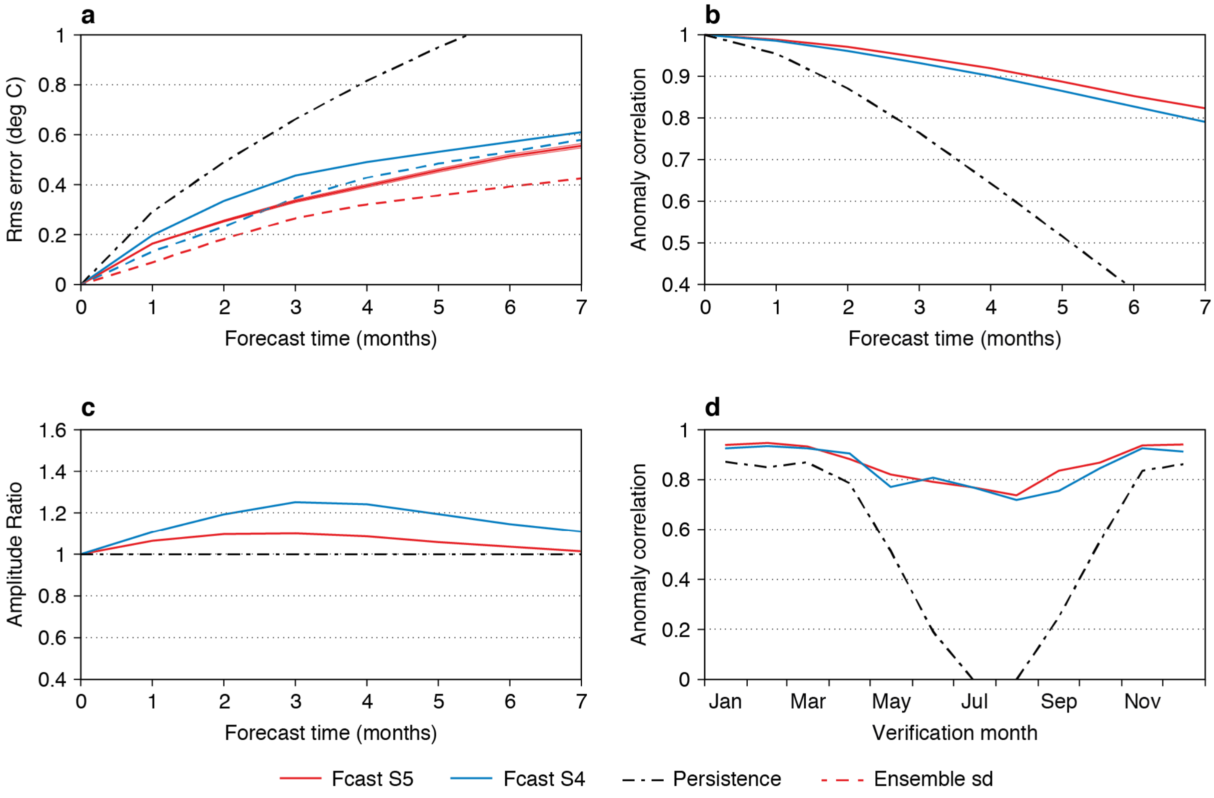


Figure 1: Forecast statistics from SEAS5 (red) and S4 (blue) for 432 cases from 1981 to 2016. (a) r.m.s. error (solid line) and ensemble spread (dashed line), (b) anomaly correlation, (c) ratio of model to observed standard deviation of anomalies and (d) anomaly correlation at 4-month lead time as a function of verification month. The dashed black lines represent an anomaly persistence forecast.

One notable feature is that the excess amplitude of SST variability seen in S4 is now substantially reduced, contributing to the improvement in RMSE. The amplitude of variability is known to be related to the model mean state, which is also much improved as discussed in section 3.2.2. The amplitude of SST variability could be corrected by calibration of the variance, as was done for S4, but this is not done in SEAS5 – see discussion in Appendix A.3. The amplitude ratio from the SEAS5 re-forecasts is made available to users of the web products as part of the verification information, and so users can in principle manually adjust the plotted Nino plumes to obtain more realistic forecast values when it is necessary to do so.

A further measure of deterministic skill is the Mean Square Skill Score (MSSS) relative to climatology. The MSSS is defined as 1-(MSE_{fc}/MSE_{clim}), where MSE_{fc} is the mean square error of the ensemble mean forecast, and MSE_{clim} the MSE of climatology. For a forecast whose ensemble members have the correct amplitude of variability, the MSSS conveys the same information as the anomaly correlation, but unlike anomaly correlation it penalizes errors in amplitude as well as phase. MSSS can also be interpreted as the fraction of variance which is correctly predicted. *Figure 2* shows the Mean Square Skill Score (MSSS) of NINO3.4 SST of the 13-month forecasts from SEAS5, run once per quarter to give an ENSO outlook. This shows a substantial improvement on S4, most of the improvement being associated with a better anomaly correlation (increasing from about 0.48 to 0.63 at month 13).

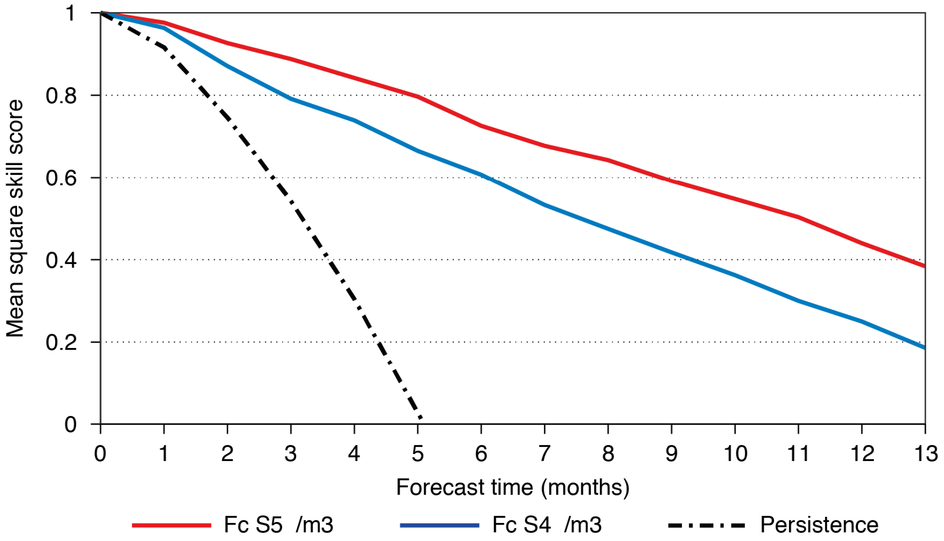


Figure 2: Mean square skill score against climatology for SEAS5 (red) and S4 (blue) forecasts of NINO3.4 SST, for forecasts up to 13 months (4 starts per year, 1981-2016).

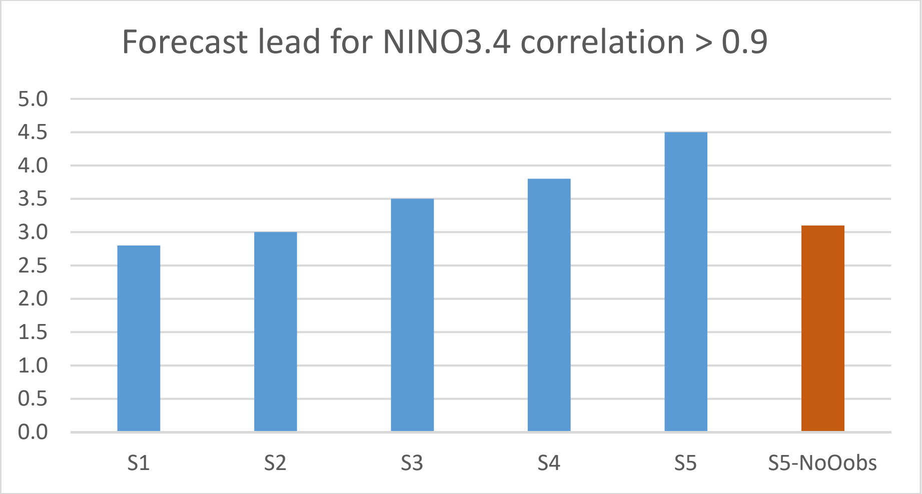


Figure 3: Lead time in months at which NINO3.4 SST anomaly correlation drops below 0.9, as measured for the five operational ECMWF seasonal forecasting systems, for the common period 1987-2002. For reference, the value for a version of SEAS5 without ocean data assimilation (SEAS5-NoOobs) is also given in orange.

We can also put the improvement in ENSO skill into a longer perspective, looking at all five real-time/operational systems that have been run at ECMWF since 1997. We compare for the available common period (1987-2002) and ensemble size (5 members). *Figure 3* shows the progress from System 1 to SEAS5 in terms of the lead time at which the NINO3.4 SST anomaly drops below 0.9. The lead time has been extended by 1.7 months over the last 21 years, a rate of improvement of about 0.8 months per decade. The increase from S4 to SEAS5 is the largest single increase to date. See *Table A 1* in Appendix A.4 for a different comparison using a metric we have presented previously to the SAC.

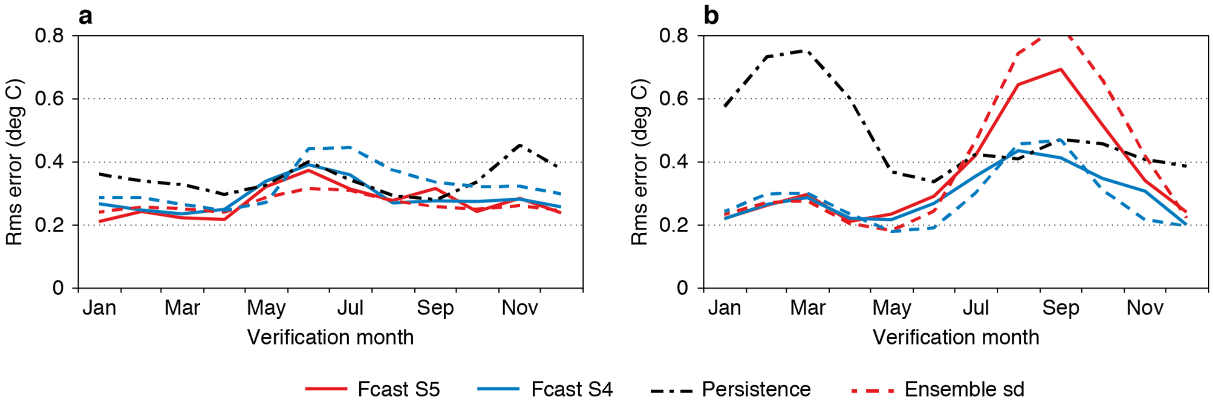


Figure 4: Comparison of SEAS5 (red) and S4 (blue) SST forecast statistics for 432 cases in 1981-2016 for (a) Equatorial Atlantic (5N-5S, 70W-30E) and (b) Eastern Indian Ocean (0-10S, 90-110E) showing r.m.s. error (solid) and ensemble spread (dashed) for forecast month 5 as a function of verification month.

Finally, we report on forecast skill for some other tropical SST regions. In the equatorial Atlantic (*Figure 4*a), where shorter-timescale ENSO like variability can occur but is predictable only at shorter leads, forecast skill is moderately improved. Ensemble spread is reduced compared to S4, and spread now matches forecast error, compared to the general over-dispersion seen in S4 for this region. In contrast with these improvements, there has been a deterioration in performance in the eastern part of the Indian Ocean Dipole (region IND2). SEAS5 produces cold events in this region (associated with a switch to easterly wind bias in NH summer) which are too large and too frequent compared to observations. In most years a substantial number of ensemble members will produce a cold event, giving a cold bias in the mean state, poor scores for the ensemble mean forecast, and a large spread in the forecast ensemble. *Figure 4*.b illustrates the impact on RMSE and spread at 4-month lead times, showing the seasonally dependent nature of the problem, and the very large ensemble spread associated with the error. This is a substantial error in the climate of the coupled forecasts, which needs to be properly understood and addressed.

3.1.2 Atmospheric forecast scores

A comprehensive set of SEAS5 seasonal forecast skill measures for all seasons, lead times and additional atmospheric variables is available online: https://www.ecmwf.int/en/forecasts/charts/catalogue/seasonal_system5_anomaly_correlation_2mtm?facets=Range,Long%20(Months)%3BType,Verification&time=2017120100,744,2018010100.

We have computed a large number of score comparisons between SEAS5 and S4 and are able to show here only a few. A starting point for assessment is to look at maps of temporal correlation between predicted (ensemble mean) and observed anomalies. Anomaly correlation is a basic measure of whether there exists some sort of correspondence between forecast and reality, and disregards how much the model output needs calibration. Appendix A.5 includes examples for 2 metre temperature and rainfall for May and November starts.

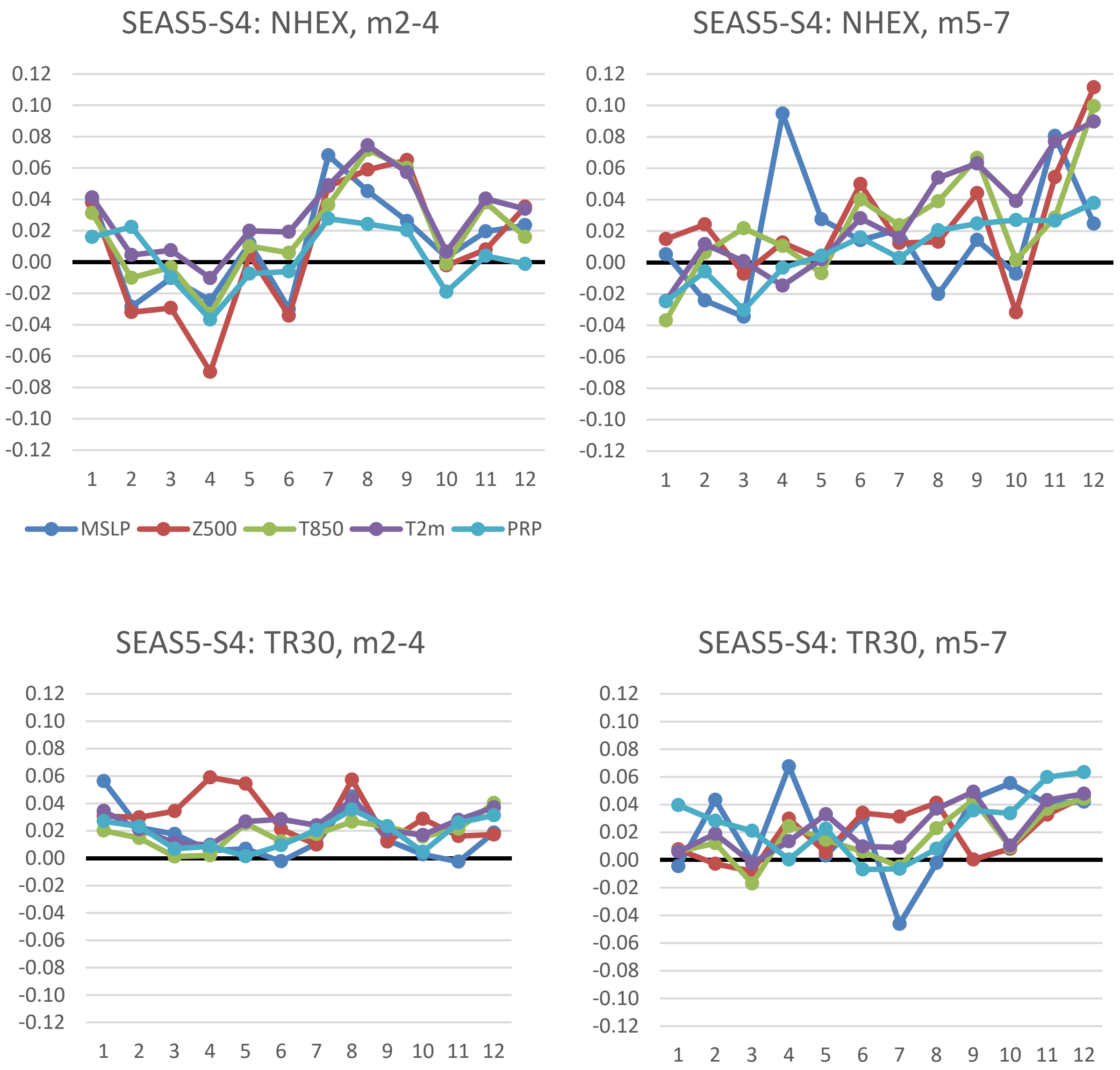


Figure 5: Spatially aggregated temporal correlation score differences between SEAS5 and S4, based on re-forecasts for 1981-2010, with 15-member ensembles and 12 start months. Coloured lines represent five different fields, score differences are plotted as a function of verification period, where 1=FMA (corresponding to Jan start dates for months 2-4) and 12=JFM (corresponding to Dec start dates for months 2-4). Top: northern extra-tropics (NHEX, 30-90N); bottom: Tropics (TR30, 30N-30S). Score differences at lead-times 2-4 months are shown on the left, and those for months 5-7 are shown on the right.

Maps of difference in correlation have the potential to show regional features but are also subject to a large amount of sampling error, especially in mid-latitudes. We can increase the statistical power of comparisons by spatially aggregating scores and by considering all start dates instead of e.g. only May and November starts. Appendix A.6 contains details of how we have aggregated scores for the tropics (TR30, 30N-30S) and northern extra-tropics (NHEX 30-90N) for five different fields (MSLP, Z500, T850, T2m and precipitation). This comparison uses only the re-forecasts of different systems (i.e. not including real-time forecasts, which have possibly inconsistent land-surface initialization), and always use identical ensemble sizes and verification periods.

Figure 5 shows the spatially aggregated temporal correlation score differences between SEAS5 and S4. A formal error analysis is given in Appendix A.6, showing that NHEX ensemble sampling uncertainty is typically in the range 0.015 to 0.020 for months 2-4, and larger at months 5-7. In any case, the consistency of the score differences can be seen directly from the plots, noting that each month is an independent sample when it comes to assessing uncertainty due to the ensembles. The tropics show consistent improvement in scores, particularly at shorter leads. The NHEX scores at months 2-4 are noisier, and seem to show a seasonal cycle with improved scores in the autumn (ASO to OND) but deterioration in late winter and spring. Longer lead NHEX forecasts are quite noisy but tend to show improved scores in late autumn and early winter.

Probabilistic scores: CRPSS

We now consider probabilistic scores, which assess how well the model ensemble performs when it is treated as representing a forecast pdf. Probabilistic scores of small ensembles perform poorly, so we compare only May and November starts, for which we have a 25-member re-forecast available for S4. We use the longest possible longer verification period (1981-2016) including the operational S4 forecasts for the period 2011-2016. This choice has a possible drawback in that there are some inconsistencies in the land surface initialization between S4 re-forecast and S4 operational run. However, comparison of scores for 1981-2010 and 1981-2016 (not shown) indicates any such effects do not alter our conclusions.

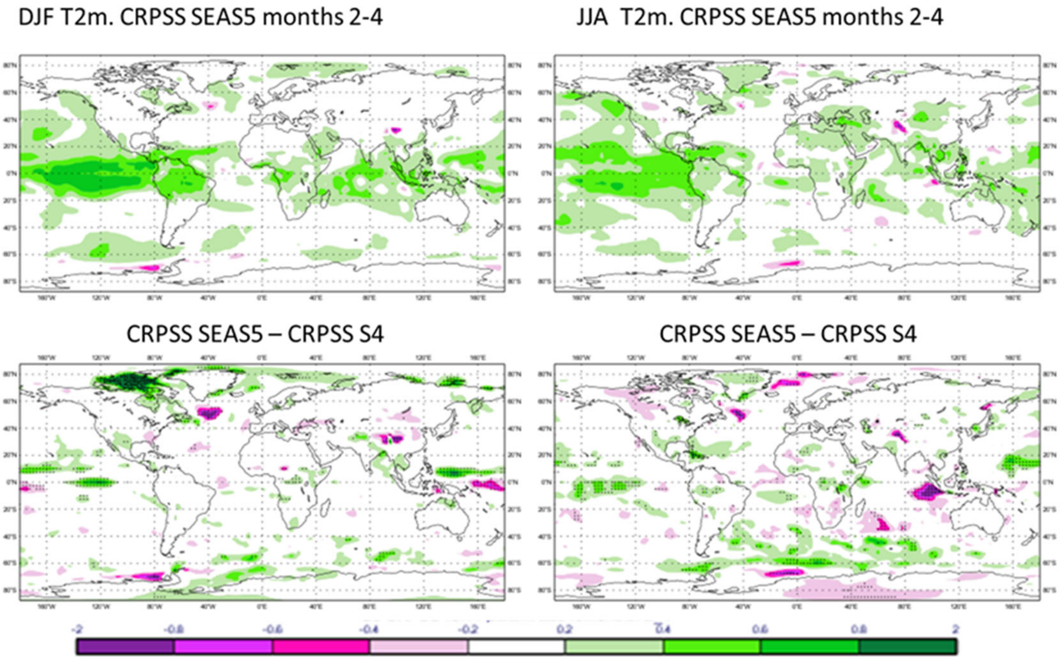


Figure 6: CRPSS of SEAS5 T2m relative to climatology, for DJF (left) and JJA (right), with differences from S4 on the lower row. Scores are calculated from 25-member ensembles over the period 1981-2016, using months 2-4 of the May/November starts. Stippling in the lower plots represents differences significant at 5%, green is where SEAS5 is better than S4.

The most appropriate overall probabilistic skill measure is the CRPS (continuous ranked probability score). The CRPS is the integral of the Brier score over all possible threshold values, for a given variable. For a deterministic forecast, the CRPS reduces to the mean absolute error. It is common to form a skill score relative to climatology, CRPSS, defined as 1-($CRPS_f$/$CRPS_{clim}$), which essentially standardizes the score to non-dimensional units, 1 being a perfect forecast and 0 being no better than climatology. Figure 6 shows spatial maps of the T2m CRPSS for months 2-4 verifying in DJF and JJA. Corresponding plots for precipitation are shown in in Appendix 0.

Some of the apparent differences between SEAS5 and S4 will be due to sampling error, which will also hide other differences which may in fact be present. Small differences are not visible due to the choice of contour intervals, but are not locally significant. Nonetheless, some differences are large and locally significant, and can be assigned to known processes and issues. The improvement in T2m scores in the sea-ice margins in winter is likely to be due to the introduction of a sea-ice model. The deterioration in scores in a region of the North Atlantic is an ocean analysis problem which is discussed in Section 3.2.7, and the problem with SST in the eastern equatorial Indian Ocean has already been discussed. We also note improvements over major lakes (Great Lakes, Caspian Sea) in JJA, related to the introduction of the lake model. Interestingly, there are also some negative signals associated with smaller lakes, such as Lake Chad in DJF and the Aral Sea in JJA. These water bodies were poorly represented in ERAI, used here for verification, and further investigation is needed to assess whether these negative signals are a cause for concern.

There are many similarities between the CRPSS difference maps and the corresponding anomaly correlation difference maps shown in Appendix A.5. There are also a few differences: some of the lake signals are not visible in the correlation differences, and more interestingly the positive CRPSS scores in the east and central Pacific are not evident in the correlation scores. Here, the CPRSS is picking up the more realistic amplitude of SST anomalies in SEAS5, which gives a more realistic distribution of actual temperature anomalies, even if there is only a marginal change in correlation.

Other forecast scores

It is possible to aggregate CRPSS scores over regions, as we did for anomaly correlation scores. This enables us to compare different forecast systems using a “scorecard” visual representation of the scores. This method of score visualization is still under development, and a recent example is shown in Appendix 0.

Reliability is another important attribute of probabilistic forecasts, and in Appendix A.8 we show some SEAS5 reliability scores. Reliability is largely unchanged from S4, with a reasonable but not perfect match between forecast and observed frequencies. Extratropical land areas are less reliable than areas over sea.

It is of interest to look at the NAO and other such indices as a diagnostic of model performance and variability. Appendix A.9 contains an analysis of NAO and PNA scores. The sampling uncertainty in NAO scores is high, and no significant difference in skill is found for either index.

3.1.3 Sea-Ice forecasts

The sea-ice cover in S4 was specified with a simple statistical model, sampling the previous five years. In SEAS5, sea-ice cover is calculated with the LIM2 sea-ice model, which gives a predictive capability but also has the potential to introduce biases.

Figure 7 shows ASO sea-ice extent in the Northern hemisphere from observations and bias-corrected forecasts started 1 July. Arctic sea ice reaches its minimum during this three-month period, which is thus of particular interest for marine applications such as ship routing and offshore operations. S4 captures the declining trend but is unable to forecast interannual variations, and the ensemble spread is problematic, being sometimes large and sometimes extremely narrow. In contrast, SEAS5 predicts both the declining trend and the interannual variations extremely well. Ensemble spread is stable and spread-error relation seems reasonable. As shown in Appendix A.10, these improved sea-ice forecasts lead to improved 2m temperature forecasts north of 70N.

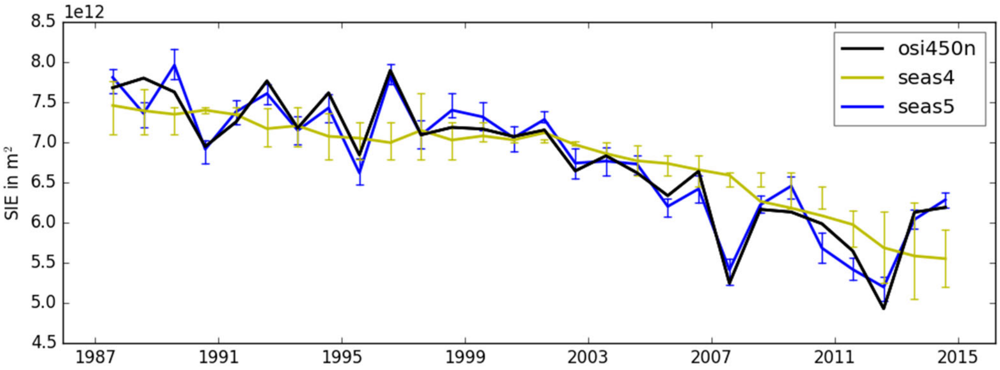


Figure 7: Northern Hemisphere sea-ice extent during the sea-ice-minimum season ASO from observations (black) and in bias-corrected forecasts started from 1st July (yellow S4, blue SEAS5). The observed sea-ice extent is calculated from monthly means of the OSI-SAF reprocessed sea-ice concentration data set available continuously from 1987 to 2014. The coloured solid lines connect the forecast ensemble means, and the error bars indicate the interquartile range of the ensemble.

Sea-ice processes are highly seasonally dependent, and the sea-ice edge resides in seas with very different local dynamics depending on the season, so model biases and forecast skill depend strongly on the target month. Furthermore, sea-ice biases develop slowly, giving a strong dependence on lead time. To form a complete picture of sea-ice forecast performance in SEAS5 in comparison with S4, we plot the difference in sea-ice area forecast error for each combination of target month and lead time. The verification data set is the OSI-SAF reprocessed sea-ice concentration, and the verification period is 1987—2014. Bias correction is applied, i.e. the mean forecast error is subtracted, before calculating the mean absolute error (MAE) of the forecasts. Since maps of differences between SEAS5 and S4 sea-ice cover MAE show distinct regional variations, we aggregate the MAE spatially to give a single number for each target month and lead time.

Figure 8(a) shows the resulting MAE difference SEAS5-S4, with periods of smaller errors in SEAS5 shaded blue. *Figure 8*(b) shows the bias which needed to be removed from SEAS5 to produce the sea-ice forecasts. For most target months and lead times, bias-corrected SEAS5 forecasts of sea ice are better than S4. Improvements are stronger for lead times of up to four months. Seasonal dependence is evident.

For target months August and September, SEAS5 is worse than S4 from lead month 4 onwards. This is related to the strong sea-ice bias in SEAS5: sea-ice does not melt enough in summer. Since the absence or presence of sea-ice comes with changed variability, this cannot be completely remedied by a-posteriori bias-correction. Likewise, forecasts of November sea-ice cover are slightly deteriorated for lead months 2-5, which again is related to a strong bias of sea ice not freezing fast enough in autumn.

Although sea-ice anomalies are better predicted by SEAS5, the summer bias more than offsets this in terms of predicting the actual ice cover. Improving the mean climate of the ice model is thus an important goal for future development. The summer and autumn biases are of opposite sign, but preliminary investigation suggests that excessive cloud cover in the Arctic might be the explanation for both, causing a lack of incoming solar radiation at the surface in summer, and slowing the rapid cooling of the surface in autumn that leads to sea-ice formation.

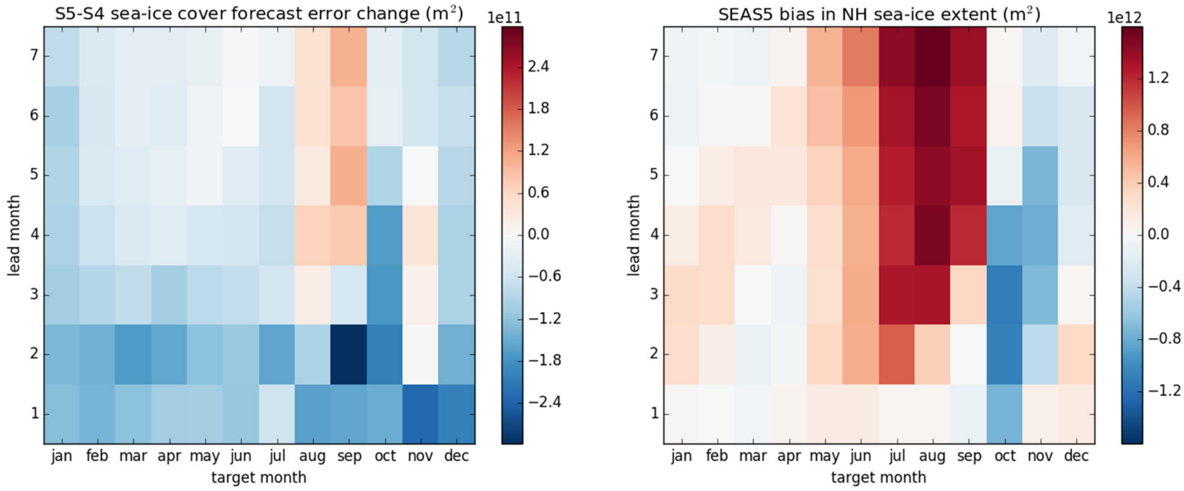
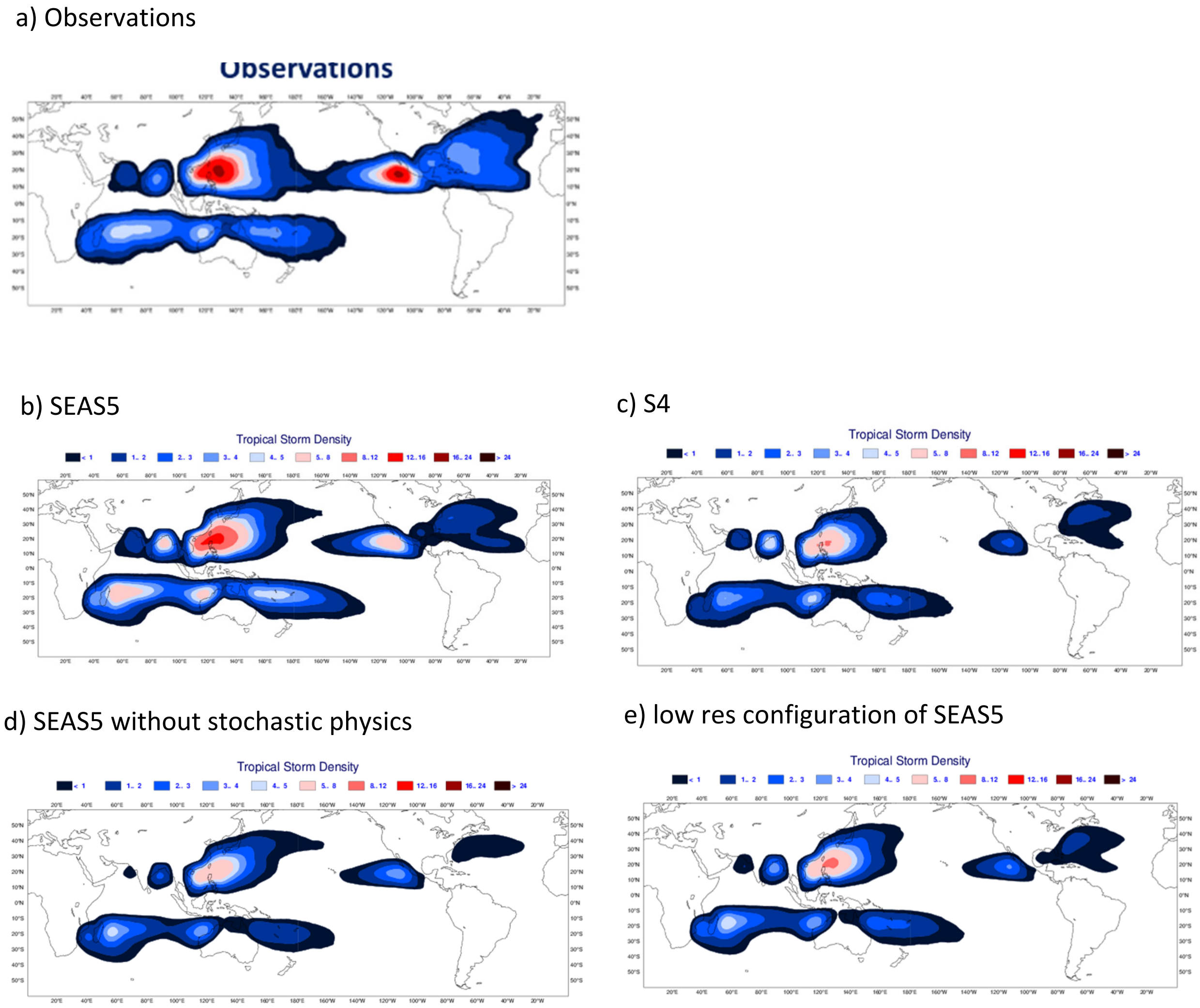


Figure 8 Left: SEAS5 - S4 spatially aggregated mean absolute error (MAE) of the bias-corrected ensemble mean sea-ice area forecast for the Northern Hemisphere, as a function of target month and forecast lead time. The magnitude of MAE change is shown by the colour bar shown on the right. Units are m^2. For reference, a change of 10^{11} m^2 corresponds to a change of sea-ice concentration forecast error of 10% in a square with 1000 km sides. Right: SEAS5 bias for Northern Hemisphere sea-ice extent with respect to ORAS5 reanalysis 1987-2014. Note the different scale.

3.1.4 Tropical cyclones

ECMWF seasonal forecasts of tropical cyclones have been issued routinely, once a month, since 2001 (Vitart and Stockdale, 2001). At the seasonal range, the tropical cyclone products include the number of tropical storms (maximum wind speed exceeding 17 m/s), number of hurricanes, accumulated cyclone energy (ACE) over several tropical cyclone basins (North Atlantic, eastern North Pacific, western North Pacific, South Indian Ocean, Australian Basin and South Pacific), tropical storm density anomaly and standardized tropical storm density for a six-month period. The tropical cyclones are detected using the tracker as described in Vitart et al *(1997)* and the statistics of detected tropical cyclones are calibrated using the seasonal re-forecasts.



a) Observations

b) SEAS5

c) S4

d) SEAS5 without stochastic physics

e) low res configuration of SEAS5

Figure 9: Tropical storm density over the period 1990-2014. The figure shows the annual number of tropical cyclones passing within 500km in observations (a), and calculated from 7-month seasonal re-forecasts initialised in May and November: b) SEAS5, c) S4, d) SEAS5 without stochastic physical parametrizations, and e) a low-resolution configuration of SEAS5 with stochastic physical parametrizations.

Figure 9 shows the climatology of tropical storm track density over the period 1990-2014 in observations (from IBTraCS https://www.ncdc.noaa.gov/ibtracs/) and various model experiments. Although still underestimating observations, SEAS5 displays a much more realistic tropical storm climatology than S4, which severely underestimated the number of tropical storms. The higher horizontal resolution of SEAS5 is the main reason for this improvement, as demonstrated by an experiment using the SEAS5 model but at a resolution close to S4 (TCo199L91/ORCA1_Z42, experiment SEAS5-lr in *Table 4*, Section 3.2.6). This shows a level of activity more like S4. Stochastic physics (SP) also plays an important role in helping the model generate a realistic climatology. When stochastic physics is switched off in SEAS5 (experiment SEAS5-noSP in *Table 4*), the tropical storm activity decreases. The impact of SP on tropical storm climatology in SEAS5 is comparable to the benefit obtained by increasing the model resolution from TCo199L91/ORCA1_Z42 to TCo319L91/ORCA025_Z75.

Table 2: Linear correlation between the interannual variability of accumulated cyclone energy over the period 1990-2014 in SEAS5 and the observed interannual variability (calculated using IBTracks for each tropical cyclone basin (atl= north Atlantic; enp=eastern North pacific, wnp=Western North Pacific, sin=South Indian Ocean, aus=Australian Basin and spc= South Pacific). Values are tabulated by the month the forecast is issued. Black numbers indicate correlations that are not significantly different to S4. Blue (red) numbers indicate correlations that are significantly larger (smaller) than in S4 (numbers in parentheses) using a 10,000 bootstrap re-sampling method. Bold numbers for SEAS5 indicate correlations that are statistically significant at the 5% level. Correlations are only calculated for the forecasts that cover a significant portion of the tropical cyclone season (for example, tropical cyclone forecasts are issued only between April and September over the North Atlantic).

	Jan	Feb	March	April	May	June	July	August	Sept.	Oct.	Nov.	Dec.
Atl	-	-	-	0.35	**0.53**	**0.64** *(0.72)*	**0.65** *(0.72)*	**0.67** *(0.62)*	**0.48** *(0.63)*	-	-	-
Enp	-	-	-	-0.09 *(0.44)*	**0.42** *(0.57)*	**0.51** *(0.67)*	0.38 *(0.44)*	0.34 *(0.56)*	0.24			
Wnp	-	-	-	**0.68** *(0.58)*	**0.70** *(0.60)*	**0.78** *(0.56)*	**0.64** *(0.70)*	**0.58**	**0.49**	-	-	-
Sin	-0.09	-0.14	0.02	-	-	-	-	-	-	0.17	0.34	0.10
Aus	-0.37	-0.24	-0.29	-	-	-	-	-	-	-0.06	-0.07	-0.29
Spc	0.14	0.28	-0.06							**0.60** *(0.47)*	0.29	**0.40** *(0.29)*

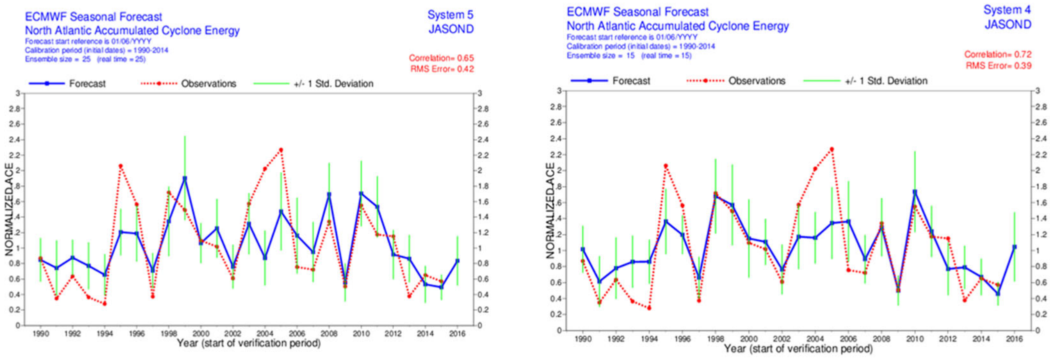


Figure 10: Inter-annual variability of Accumulated Cyclone Energy (ACE) over the North Atlantic from July to December 1990 and 2014 in SEAS5 (blue line, left panel), S4 (blue line, right panel) and Observations (red line). The vertical green lines represent 2 standard deviations. The forecast start date is 1st June.

The improvement in tropical storm climatology does not necessarily translate into more skilful forecasts of tropical cyclone inter-annual variability. *Table 2* shows the linear correlation between the interannual variability of tropical cyclone accumulated cyclone energy (ACE) in SEAS5 and observations from IBTrACS. SEAS5 displays significant skill in predicting the interannual variability of tropical cyclone ACE over the North Atlantic, eastern North Pacific (May and June start dates only), western North Pacific and South Pacific (October and December start dates). SEAS5 displays generally lower skill than S4 over the Atlantic and eastern North Pacific but higher skill over the western North Pacific (particularly from April to June) and over the South Pacific. Detailed examination of time-series of SEAS5 forecasts over the North Atlantic (*Figure 10*) suggests that the deterioration is modest, and perhaps driven by small changes in a few individual years (1998, 1999, 2004). Although the slight drop in skill is not fully understood, it does not seem to represent any major change in forecast characteristics.

3.2 Understanding Aspects of Performance

Having outlined the seasonal forecast performance of SEAS5, we now examine various aspects of SEAS5 which help us understand the forecast performance. We will make use of several area average indices, some of which are well known, but some of which are specific to our analyses. *Table 3* provides the coordinates of the areas used.

A realistic representation of the tropical climate, both mean and variability, is a crucial requirement for a successful seasonal forecasting system. The model climate is also relatively easy to sample, and should be a starting point in assessing any seasonal forecasting system. We first describe the changes in climate with successive IFS cycles, and then at look at some aspects of the SEAS5 climate in more detail.

Table 3: Definition of area average indices

NINO3.4	5N-5S	120-170W	**IND2**	0-10S	90-110E
NINO3	5N-5S	90-150W	**WCIO**	10N-10S	40-90E
NINO4	5N-5S	160E-150W	**NWATL**	30-70N	100-40W
NINO4W	10N-10S	160E-150W	**NASD**	45-55N	50-30W
EQ3	5N-5S	150E-170W			

We have also undertaken a number of experiments with variants of SEAS5, to discuss dependencies on specific aspects of the SEAS5 configuration. These are listed in *Table 4* in Section 3.2.6.

3.2.1 Evolution of model climate between S4 and SEAS5

The climatology of the seasonal forecasts depends on both how the system is configured (resolution, stochastic physics, other choices) and on the model cycle used. Between S4 and SEAS5 there were 8 upgrades of the IFS atmospheric model and 1 update to the ocean model. A summary of the upgrades is listed in Appendix A.11.

The model climate has been evaluated for each model version between S4 (cycle 36r4) and SEAS5 (cycle 43r1) with a fixed experimental setup based on the S4 model configuration. Simulations have been carried out for start dates of 1 May and 1 November from 1981 to 2010 with 7-month integrations and either 3 (up to cycle 38r2) or 10 ensemble members. The horizontal and vertical resolution has been kept the same as S4, and stochastic physics has been switched off. From the November start dates DJF and MAM statistics are aggregated and from May start dates the statistics for JJA and SON. Simulations have been carried out both in coupled mode (ORCA1_Z42, no dynamical sea-ice model, as in S4), and in uncoupled mode by relaxing the SSTs to observed values.

A brief summary of some findings from these runs is given in Appendix A.12, tracking the stages by which model biases in mid-tropospheric temperature have been reduced. Overall, there have been improvements in global metrics for model climate, and substantial improvement in the tropical circulation, with an overall reduction of the cold tropical bias (see sections 3.2.2 and 3.2.3). Not all changes in model climate have been beneficial, and we show here one field whose climate has deteriorated. Unlike later analysis, which compares the SEAS5 and S4 climate, here we are looking at the effect of IFS Cycle changes only, with all other aspects of the model configuration fixed.

Upper level winds are important for teleconnections and a useful diagnostic for the model circulation. *Figure 11* shows 200hPa zonal wind bias from S4 and Cy43r1, indicating a deterioration in the winds, in particular for the location and strength of the sub-tropical/mid-latitude jet, especially on the summer hemisphere. This is a hard problem to tackle, since it involves a delicate balance between temperature gradients and momentum balance. We plan to devote efforts to improve this error in future model cycles: the higher-resolution of SEAS5 does not solve this problem. The zonal mean structures associated with it are described in the next section.

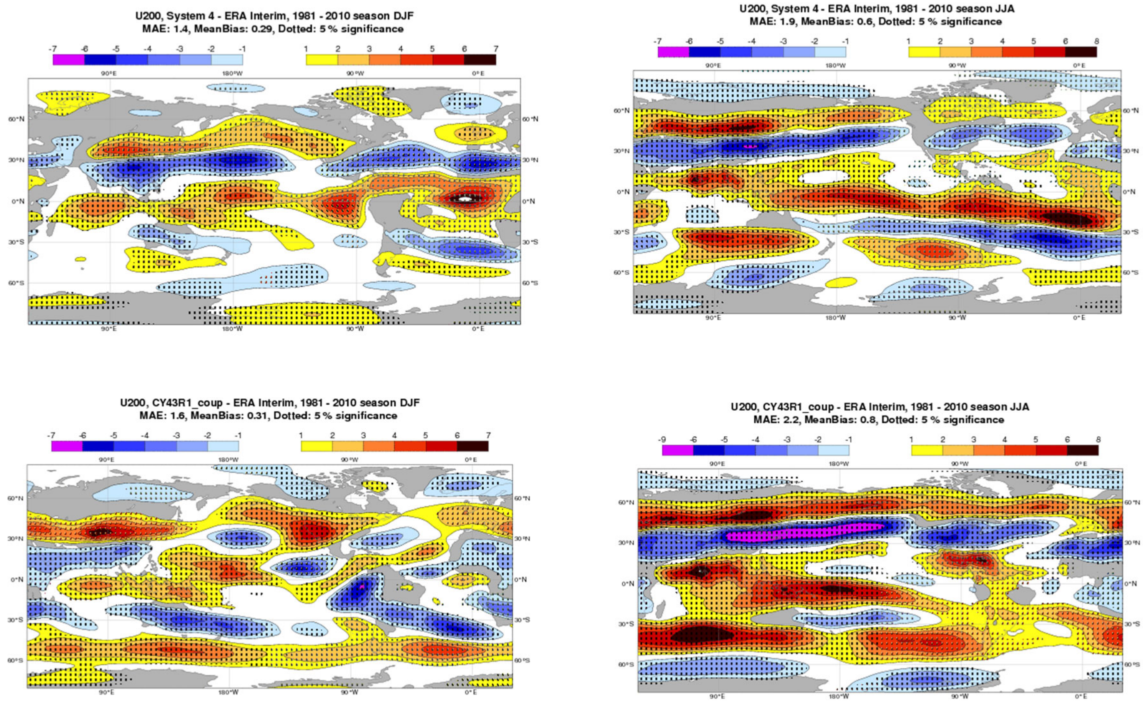

U200, System 4 - ERA Interim, 1981 - 2010 season DJF
MAE: 1.4, MeanBias: 0.29, Dotted: 5 % significance
U200, System 4 - ERA Interim, 1981 - 2010 season JJA
MAE: 1.9, MeanBias: 0.6, Dotted: 5 % significance
U200, CY43R1_coup - ERA Interim, 1981 - 2010 season DJF
MAE: 1.6, MeanBias: 0.31, Dotted: 5 % significance
U200, CY43R1_coup - ERA Interim, 1981 - 2010 season JJA
MAE: 2.2, MeanBias: 0.8, Dotted: 5 % significance

Figure 11: 200 hPa zonal wind bias with respect to ERA Interim for S4 (top), and cycle 43r1 (bottom) for DJF (left) and JJA (right). DJF on the left and JJA on the right.

The evaluation of model climate involves a large number of diagnostics calculated and plotted by a processing suite. The aim is to give feedback and guidance to the model developers. One limitation of this monitoring process is that it runs the model without stochastic physics: this is in line with the deterministic model development work at ECMWF, but gives a disconnect to the actual physics used in medium-range, extended and seasonal ensemble forecasts. The effects on the model climate at seasonal time-scales of different stochastic schemes were evaluated in Leutbecher et al (2017). Weisheimer et al (2014) and Subramanian et al (2017) also document the impact of SPPT on S4. For the model climate, the largest change in stochastic physics between S4 and SEAS5 was the introduction of a global conservation fix in the SPPT scheme in model cycle 43r1, based on work done within the EC-Earth Consortium. We plan to update the seasonal climate test suite to include stochastic physics, along with other updates to the preferred low-resolution configuration (details in section 3.2.6).

A second limitation of the climate evaluation is that the plethora of diagnostics produced can make it hard to focus on the key aspects of climate that are most critical for seasonal prediction. For example, there was not much awareness of the deterioration in 200hPa winds until the process of building SEAS5 was underway. It is planned to complement the current evaluation of model climate for each model cycle with a more comprehensive evaluation of seasonal skill scores with a configuration based on SEAS5. Details of this configuration are being planned, along with how the diagnostic output can best be organized and presented.

3.2.2 SEAS5 mean state

We now examine some aspects of the forecast mean state in SEAS5 itself. *Figure 12* shows the SST bias in S4 and SEAS5, relative to the ocean reanalysis they were initialised from, ORAS4 (Balmaseda et al, 2013) or ORAS5 (Zuo et al, 2018). The tropical oceans are warmer in SEAS5, especially in the summer hemisphere, and give an overall positive bias to tropical SST. Warmer biases flank the equator in the tropical Pacific and Atlantic basins. In the Indian Ocean and west Pacific, cold biases in S4 are replaced with a warm bias in SEAS5. Importantly, SEAS5 gives a major reduction in the equatorial Pacific cold tongue bias, which was one of the biggest problems in S4.

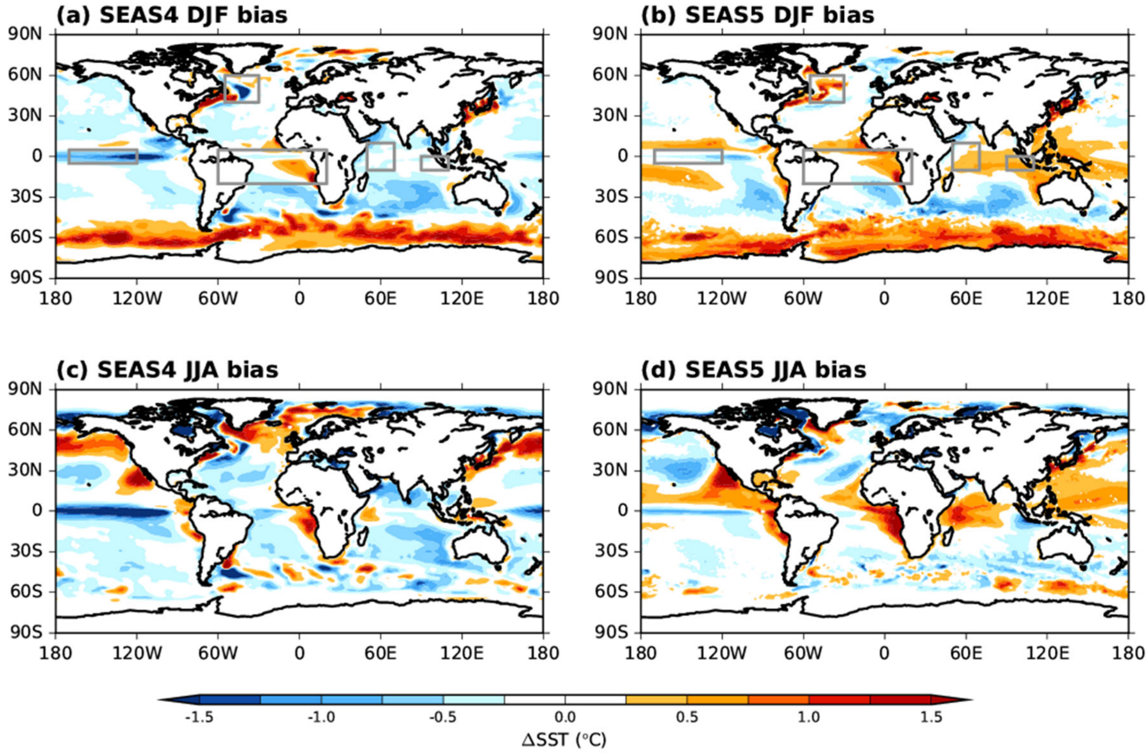


Figure 12: Winter and summer SST bias for forecast lead 2-4 months for S4 (a,c) and SEAS5 (b,d) relative to the SST from ORAS4 and ORAS5 respectively

The reduction of the cold tongue bias is of dynamical origin, a consequence of increased ocean resolution and improved equatorial winds in the 43r1 model cycle. The large-scale reduction of tropical cold biases in SEAS5 appears related to changes in the atmosphere radiative balance: improvements in the IFS model physics for tropical convection and clouds gives higher total column water vapour in SEAS5, with more absorption of thermal radiation, resulting in a reduction in tropical outgoing long-wave radiation. In the northern Pacific SST biases also reduce, particularly in the summer. This is due, at least in part, to improved parametrizations for ocean vertical mixing. Changes in the North Atlantic are discussed in Sections 3.2.6 and 3.2.7.

To examine changes in the atmospheric mean state, we show in *Figure 13* the zonally averaged temperature profile bias with respect to ERA-Interim in S4 and SEAS5 for both DJF and JJA. The zonal wind profile bias is over-plotted as contours. The model troposphere is warmer in SEAS5 than S4, a clear decrease in bias both in DJF and JJA. The SEAS5 troposphere is however now slightly too warm in JJA. Some temperature gradients are not as well represented in SEAS5 as they were in S4 – this is especially true at the 250-150hPa level, due to the strong contrast between tropical tropospheric biases (slightly warm) and mid-latitude lower stratospheric biases (very cold). In JJA there is also a tropospheric warming from approximately 30◦N to 40◦N. The SEAS5 jets are too strong at the tropopause level in both seasons, but in JJA errors extend lower and the jets are positioned too far to the north in both hemispheres.

The lower stratospheric cold biases worsen in SEAS5 in part due to the increase in horizontal resolution which increases the level of resolved vertically propagating gravity wave activity (Polichtchouk et al, 2017), and in part due to long-standing humidity errors in the lower stratosphere (Hogan et al, 2017). It is unclear whether the JJA temperature and wind errors at 40N in the mid-troposphere have a tropospheric origin, or are forced from the lower stratosphere/tropopause errors above. The increased biases in zonal mean winds at tropopause level are a potentially serious deterioration of SEAS5 compared to S4, due to their importance for dynamical forcing of the extratropics by the tropics. Recent targeted experimentation carried out within the stratospheric task force has identified a few directions for model development that should lead to bias reduction in the upper troposphere and stratosphere region (Shepherd et al, TM824, Hogan et al, TM816).

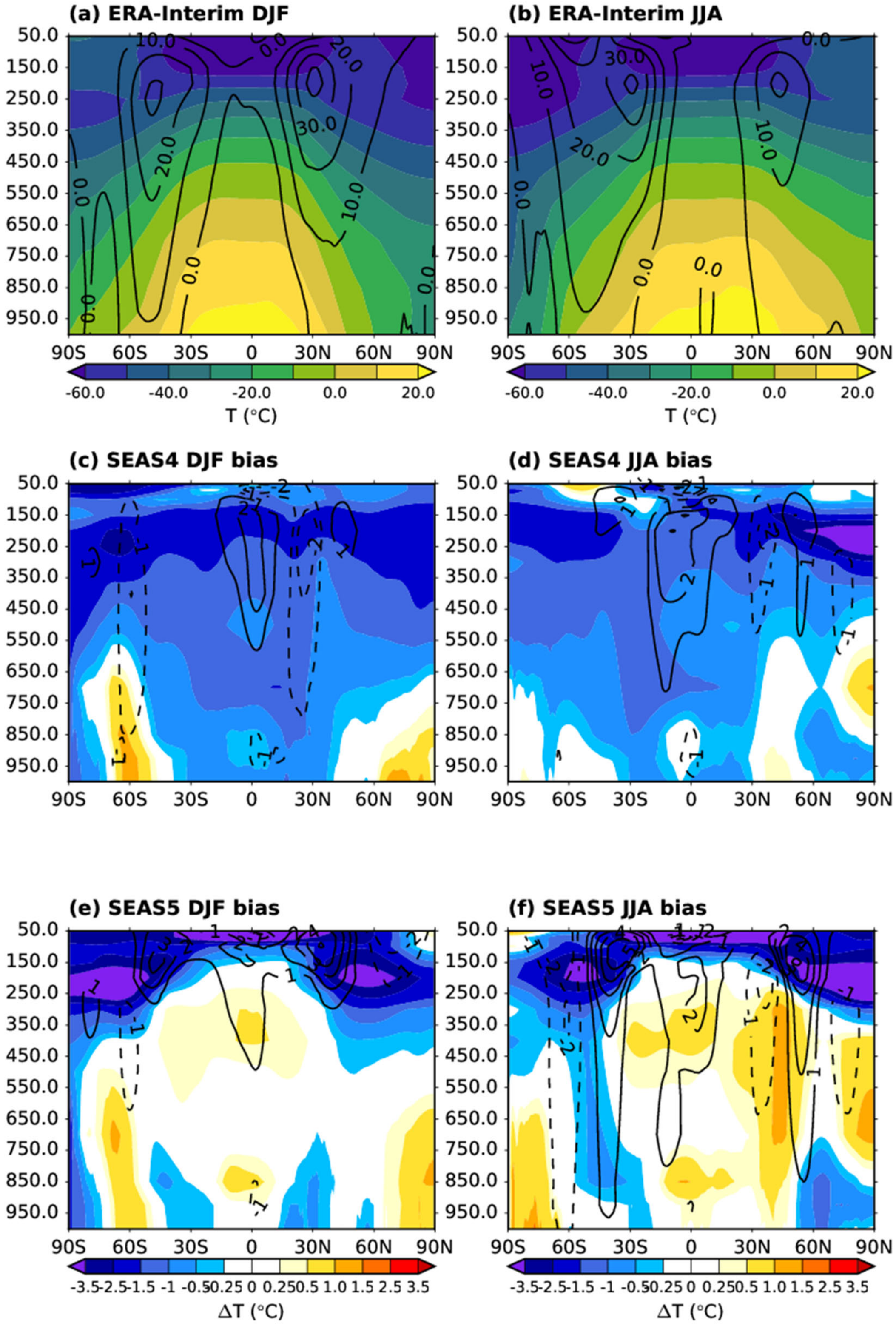


Figure 13: Zonal mean sections of bias relative to ERA-Interim for DJF (left) and JJA (right), for S4 (top row) and SEAS5 (bottom row). Temperature bias is indicated in colour, and zonal wind bias with contours. Biases are for months 2-4.

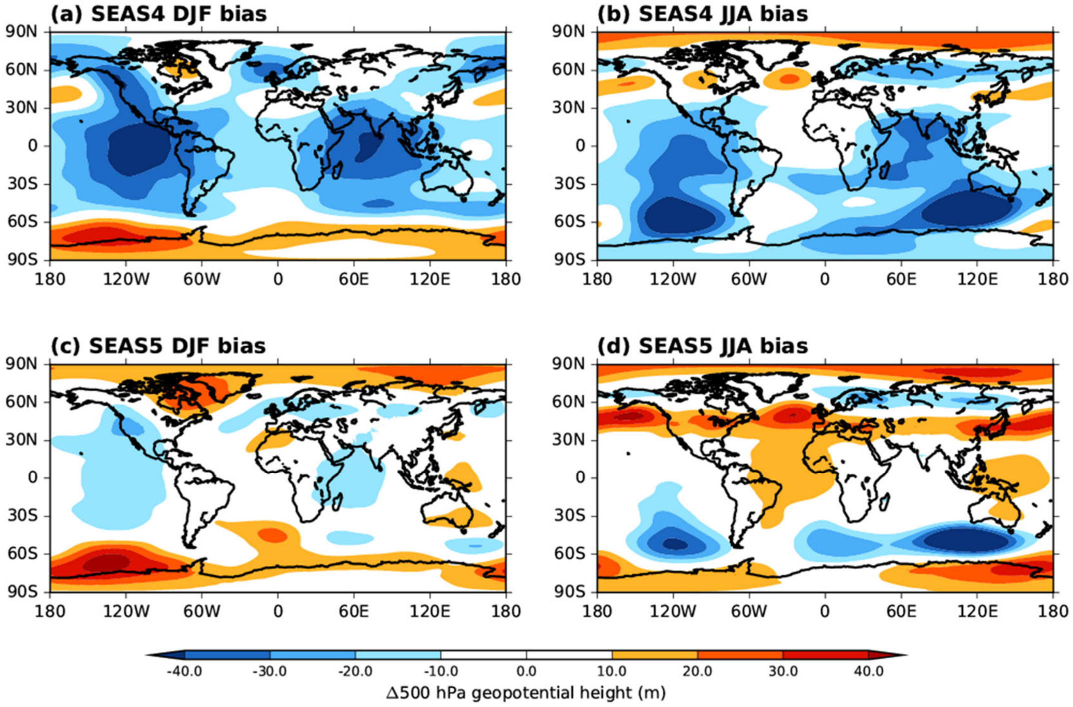


Figure 14: Winter and summer 500 hPa geopotential height bias in S4 (a,b) and SEAS5 (c,d) with respect to ERAI, for forecast lead-times 2-4 month. Shown are forecast initialized in November (left) and May (right).

Figure 14 shows the horizontal structure of SEAS5 biases in Z500 relative to ERA-Interim. The warming of the troposphere in SEAS5 is reflected in higher geopotential heights, and this substantially reduces the bias both in winter and summer. In summer however, the displacement of the jet is clearly visible and enhanced in SEAS5 compared with S4. Equivalent biases for MSLP are shown in Appendix A.13, together with an analysis of blocking, which shows a minor improvement in the Pacific but no change in the Atlantic.

3.2.3 Air-sea interaction in the tropics

Here we discuss low-level tropical wind circulation and precipitation, with a focus on the central Equatorial Pacific (region NINO3.4) and Eastern Indian Ocean (IND2), where SEAS5 and S4 differ in their skill (*Figure 6*, Section 3.1.2).

Figure 15 shows SEAS5 biases with respect to ERA-I in zonal wind at 850 hPa (U850) for lead times 2-4 in forecast initialized in November (verifying in DJF, left) and in May (verifying in JJA, right). The bottom panels show the equivalent differences between SEAS5 and S4. The persistent easterly bias in Equatorial Pacific west of the date line is still present in SEAS5, especially in May starts, although it has been substantially reduced w.r.t. S4. The easterly bias in S4 was very severe, enhanced by the positive coupled Bjerknes feedback (Molteni et al, 2011). This can be seen in *Figure 16*a, which summarizes the zonal wind bias at 4-month lead time over a region EQ3 in the Western Equatorial Pacific in S4 and SEAS5 (May starts). In SEAS5 the reduction of the easterly bias stems from the combined effect of the atmospheric model cycle and resolution. A positive Bjerknes feedback is still present in SEAS5 mean errors: the same atmospheric model as SEAS5 forced by observed SST exhibits an easterly bias about 30% weaker than in coupled mode.

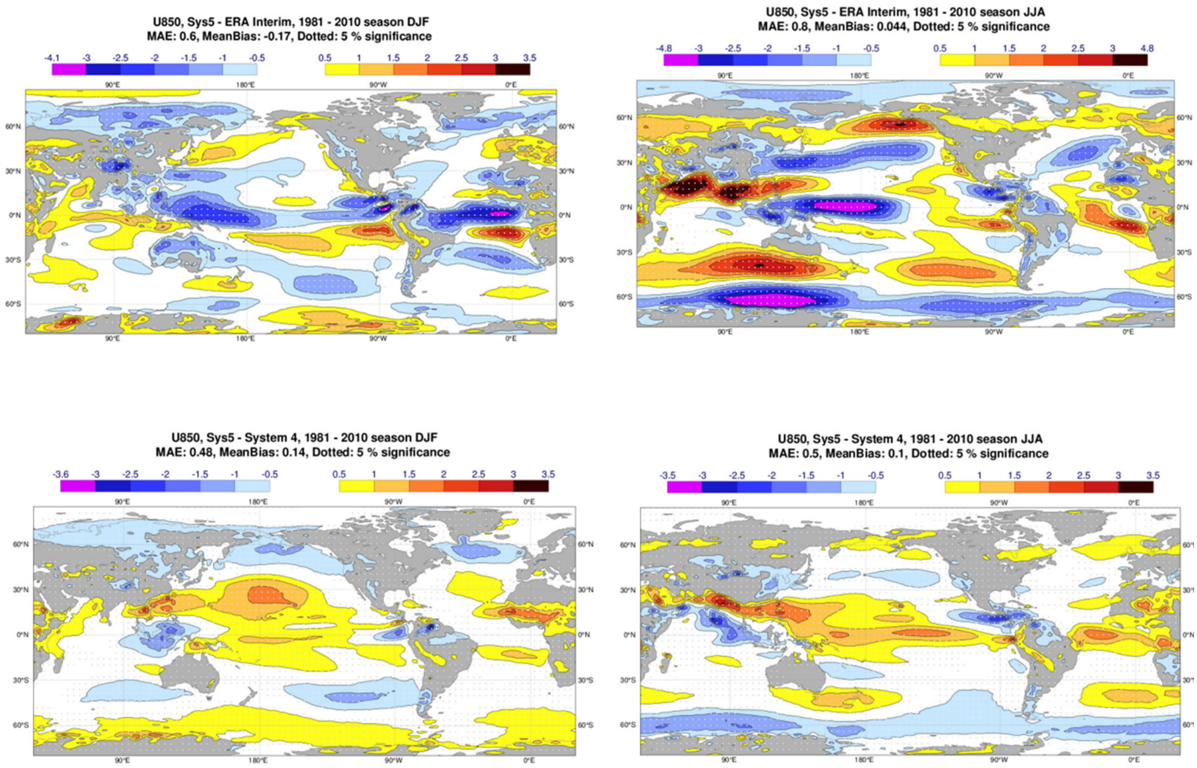

U850, Sys5 - ERA Interim, 1981 - 2010 season DJF
MAE: 0.6, MeanBias: -0.17, Dotted: 5 % significance
-4.1 -3 -2.5 -2 -1.5 -1 -0.5 0.5 1 1.5 2 2.5 3 3.5
U850, Sys5 - ERA Interim, 1981 - 2010 season JJA
MAE: 0.8, MeanBias: 0.044, Dotted: 5 % significance
-4.8 -3 -2.5 -2 -1.5 -1 -0.5 0.5 1 1.5 2 2.5 3 4.8
U850, Sys5 - System 4, 1981 - 2010 season DJF
MAE: 0.48, MeanBias: 0.14, Dotted: 5 % significance
-3.6 -3 -2.5 -2 -1.5 -1 -0.5 0.5 1 1.5 2 2.5 3 3.5
U850, Sys5 - System 4, 1981 - 2010 season JJA
MAE: 0.5, MeanBias: 0.1, Dotted: 5 % significance
-3.5 -3 -2.5 -2 -1.5 -1 -0.5 0.5 1 1.5 2 2.5 3 3.5
90°E
180°E
90°W
0°E
60°N
30°N
0°N
30°S
60°S

Figure 15: Top panels: SEAS5 biases in U850 with respect to Era-I for months 2-4 in forecast initialized in November (left) and May (right). The bottom panels show the differences in the mean U850 between SEAS5 and S4.

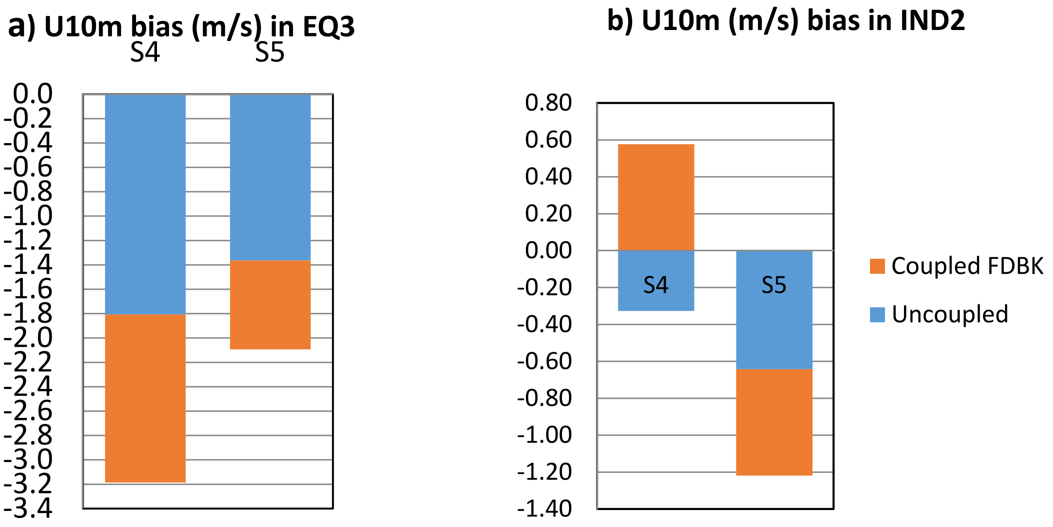


Figure 16: Bias in U10m (m/s) over the Western Equatorial Pacific (region EQ3, left) and Eastern Indian Ocean (region IND2, right) in S4 and SEAS5 forecast initialized in May. The blue bars show the bias in the uncoupled seasonal integrations forced by observed SST. The orange component is the differences between the coupled and uncoupled integrations. The coupling enhances the wind biases, except in S4 in region IND2.

Over the Indian Ocean, wind biases in SEAS5 are quite striking. Over the Eastern Equatorial IND2 region, SEAS5 wind bias during JJA is easterly, and it generates a positive feedback (*Figure 16*a), which almost doubles the easterly bias, leading to an unrealistic cold-tongue regime in this region, with a cold and dry bias (precipitation biases shown in Appendix A.14). During this season SEAS5 overestimates the interannual variability, the forecasts have a large spread and SEAS5 does not beat a persistence forecast, as shown in *Figure 4*b. Curiously in S4 the wind bias over this region was quite small and positive (*Figure 16*b); the uncoupled model had a small negative bias, but the coupling produced a feedback in the opposite direction. North of the Equator, the summer Monsoon in SEAS5 has too strong a zonal component, extending far too much into the Philippine sea (*Figure 15* b, d). This is a degradation with respect to S4, and is stronger in the uncoupled model.

Further aspects of the tropical wind and precipitation biases are discussed in Appendix A.14. Overall, the circulation in SEAS5 is too zonal and too symmetric around the Equator, likely due to errors in the atmospheric heating and surface drag (Simpson et al, 2018). These errors are enhanced in the coupled model, with the Pacific and the Maritime Continent as the main areas where the coupling enhances the wind biases and changes the patterns of precipitation.

3.2.4 Tropical-Extratropical Teleconnections

Teleconnections from the tropics are an important source of predictable signals for the extratropical regions. Although they can be detected throughout the whole yearly cycle, many teleconnection patterns affecting the northern midlatitudes reach their largest amplitude during the northern winter, when the strong vorticity gradients in the subtropical regions intensify the Rossby wave sources associated with tropical convection (Sardeshmukh and Hoskins, 1988).

A detailed analysis of teleconnections originated from tropical Indo-Pacific rainfall anomalies during the northern winter in the ECMWF seasonal S4 was carried out by Molteni et al (2015*,* MSV15 hereafter). Overall, S4 provided a good simulation of the relationship between SST and rainfall anomalies within the tropical belt, and of extratropical teleconnections to the North Pacific – North American sectors. On the other hand, teleconnections to the Euro-Atlantic sector in S4 showed significant differences from the corresponding observed patterns, with an underestimation of the link between western/central Indian Ocean rainfall and NAO variability, and an incorrect phase of the ENSO response over the North Atlantic (see Fig. 6 in MSV15). The latter problem was linked to an excessively strong correlation between rainfall anomalies in the Nino4 region and the western/central Indian Ocean (WCIO).

Although a more detailed analysis of teleconnections in SEAS5 will be provided elsewhere, here we can summarize our preliminary analysis as follows:

- Connections between tropical SST and rainfall show relatively minor changes with respect to S4; this implies an overall satisfactory performance, but also the persistence of the too-strong correlation between NINO4W and WCIO rainfall (see *Figure 17*).
- Teleconnections into the Euro-Atlantic sector show bigger differences from S4, with an improved pattern associated with central Pacific anomalies, but a substantial failure in reproducing the NAO connection with WCIO rainfall (see *Figure 18*).

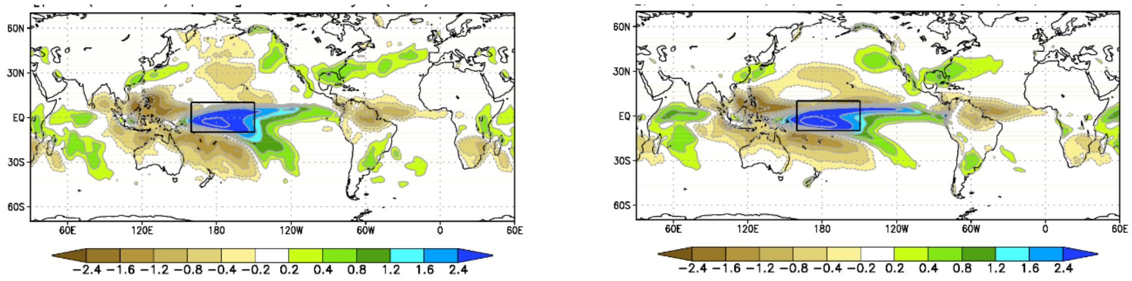


Figure 17: Covariance between normalised DJF rainfall anomaly in the NINO4W region (black box) and rainfall anomaly elsewhere. Left: GPCP v2.3 data from DJF 1981/82 to 2016/17; right: from SEAS5 re-forecasts started on 1 Nov 1981 to 2016. Note the stronger signal over the western Indian Ocean in SEAS5.

The reasons for both the improvements and deteriorations of extratropical teleconnections in SEAS5 are still being investigated. The improved simulation of the ENSO response is consistent with the general improvements in the representation of ENSO reported in previous sections of this paper. With regard to the deterioration of the WCIO-North Atlantic connection, it is relevant to note that a set of re-forecasts run with the same atmosphere-land configuration and initial conditions as SEAS5 but with prescribed, observed SST show a teleconnection pattern in much better agreement with observation (top-right panel in *Figure 18*). Also, hindcast experiments with SEAS5-lr (see *Table 4*) show a more realistic teleconnection between WCIO and the North Atlantic than SEAS5 (not shown). A poor simulation of the teleconnection pattern from Indian Ocean rainfall has also been noticed in multi-decadal coupled simulations performed with the same IFS and NEMO versions used in SEAS5 (see Roberts et al, 2018 for an overview on these experiments), while similarly long simulations with prescribed, observed SST showed a much better agreement with observations (see *Figure 19*).

Since links between Indian Ocean rainfall and the NAO are also evident on the sub-seasonal time scale (Cassou, 2008; Lin et al, 2009), it is likely that deficiencies in the SEAS5 performance in reproducing tropical intra-seasonal variability (such as the Madden-Julian Oscillation, MJO) and the associated ocean-atmosphere feedbacks may have a common cause with the teleconnection errors detected on the seasonal scale (see the significant decrease in the occurrence of MJO phases with active convection over the Indian Ocean, diagnosed in the next section).

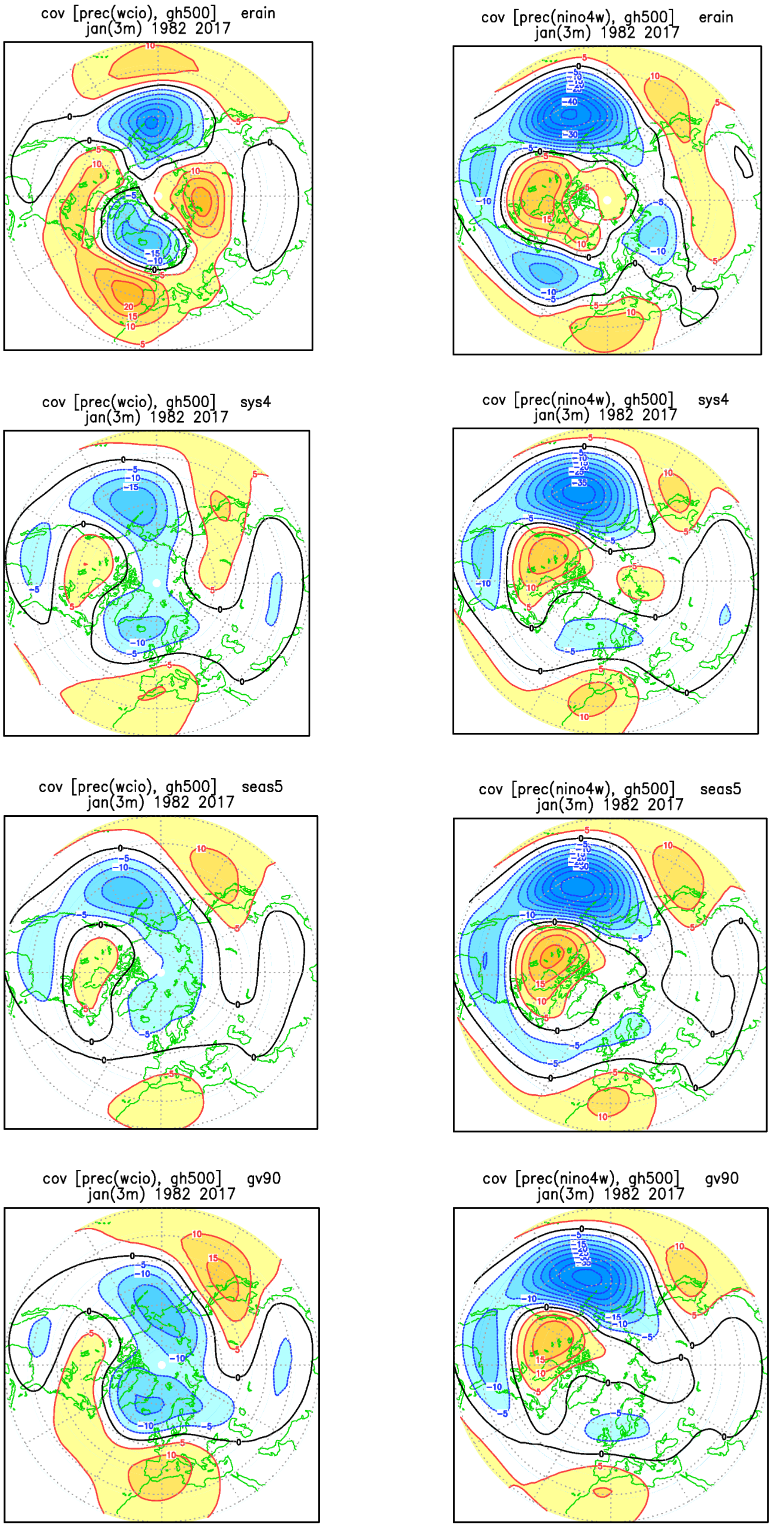

cov [prec(wcio), gh500] erain
jan(3m) 1982 2017
cov [prec(nino4w), gh500] erain
jan(3m) 1982 2017
cov [prec(wcio), gh500] sys4
jan(3m) 1982 2017
cov [prec(nino4w), gh500] sys4
jan(3m) 1982 2017
cov [prec(wcio), gh500] seas5
jan(3m) 1982 2017
cov [prec(nino4w), gh500] seas5
jan(3m) 1982 2017
cov [prec(wcio), gh500] gv90
jan(3m) 1982 2017
cov [prec(nino4w), gh500] gv90
jan(3m) 1982 2017

Figure 18: Covariances between normalised DJF rainfall anomalies in the western/central Indian Ocean (WCIO, left) and NINO4W (right) regions, and 500hPa height anomalies over the northern extratropics. Top row: from GPCP v2.3 rainfall and ERA-interim height data in DJF 1981/82 to 2016/17; second row: from S4 re-forecasts started on 1 Nov 1981 to 2016. (cf. Fig. 6 in MVF15); third row: as above, but from SEAS5 re-forecasts; bottom row: as above, but from SEAS5 ensembles with prescribed, observed SST. All ensembles include 25 members.

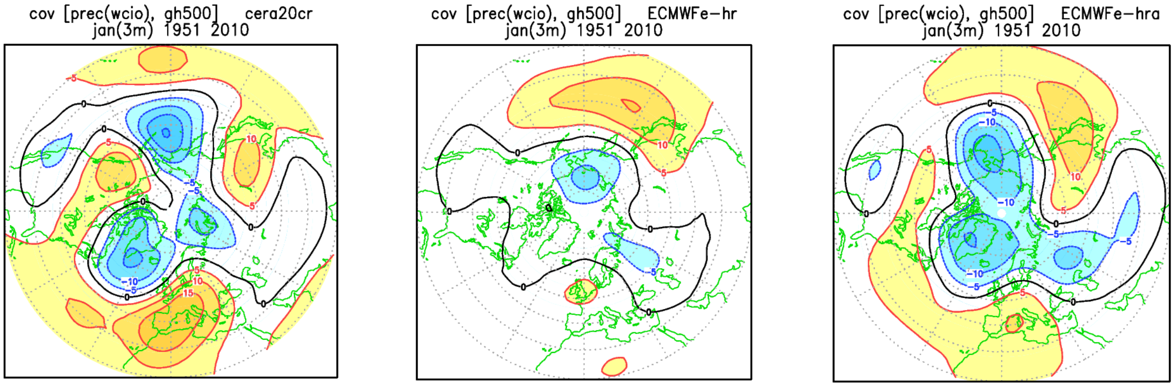

cov [prec(wcio), gh500] cera20cr
jan(3m) 1951 2010
cov [prec(wcio), gh500] ECMWFe-hr
jan(3m) 1951 2010
cov [prec(wcio), gh500] ECMWFe-hra
jan(3m) 1951 2010

Figure 19: Covariances between normalised DJF rainfall anomalies in the western/central Indian Ocean (WCIO) and 500-hPa height anomalies over the northern extratropics in 60 winters, from DJF 1950/51 to DJF 2009/10. Left panel: from CERA20C data; central panel: from the coupled historical simulations run with the IFS-NEMO model for the EU-H2020 PRIMAVERA project; right panel: from the PRIMAVERA historical simulations with prescribed SST from HadISST2 (see Roberts et al, 2018).

SEAS5 tropical-extratropical teleconnections will be further analysed as part of an ongoing intercomparison project organized by WGSIP (Working Group on Seasonal to Interdecadal Predictions). In the first part of this project, Scaife et al (2018) show that predictions of tropical rainfall alone can generate highly skilful forecasts of the main modes of extratropical circulation via linear relationships that might provide a useful tool to interpret real time forecasts.

3.2.5 Madden Julian Oscillation

The Madden Julian Oscillation has been diagnosed in the SEAS5 25-member ensemble re-forecasts using the Wheeler and Hendon index (Wheeler and Hendon, 2003). Since the Madden Julian oscillation is particularly active during winter and spring, the evaluation is for the period January to March for lead times from 1 to 6 months. *Figure 20* shows the evolution of the normalized MJO amplitude of the forecast relative to ERA Interim as a function of lead time for the verification period January to March. The mean amplitude of the MJO in month 1 is about 10% weaker than in ERA Interim, and this underestimation remains stable during the 6 months of integration, with values between 10% and 15%. S4 also underestimated the MJO amplitude (right), but less so than SEAS5. Deactivating the stochastic physics in SEAS5 results in a further 10% reduction of the MJO activity and a corresponding decrease in MJO spread (see Appendix A.15) in line with the findings reported by Weisheimer et al, 2014 for S4. The MJO is an important source of variability in the extratropics at the sub-seasonal timescales through in particular its impact on the North Atlantic Oscillation (NAO) (see for example Cassou, 2008), and the underestimation of the MJO amplitude may impact NAO variability in SEAS5. It is worth noticing that the amplitude of the MJO has increased in the recent IFS cycle Cy45r1.

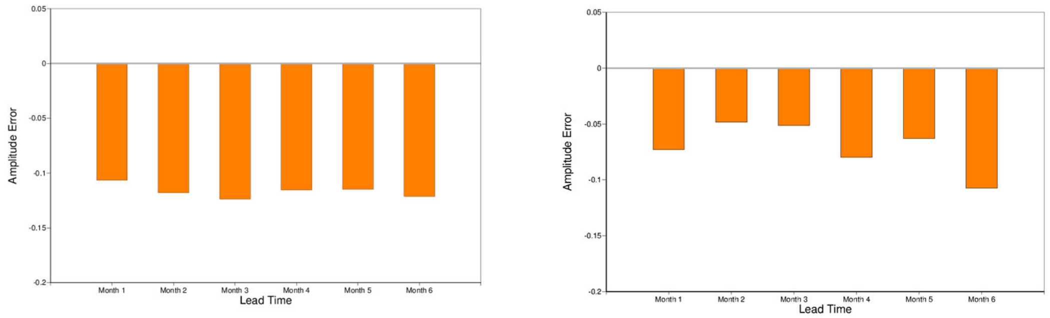


Figure 20: MJO amplitude ratio (forecast versus ERA Interim) as a function of lead time (months 1 to 6) for SEAS5 (left) and S4 (right)

Another important aspect of the MJO is its propagation from the Indian ocean to the western Hemisphere. *Figure 21* shows the ratio of days when a strong MJO is active over the Indian Ocean, the Maritime Continent, the western Pacific or the western Hemisphere. For SEAS5, in months 1-3 the statistics are close to ERA interim, and more realistic than S4. However, as the forecast lead time increases, the number of days with an MJO over the Indian Ocean and the western hemisphere diminishes, while the MJO becomes more frequent over the western Pacific and the Maritime Continent after a lead time of 4 months. There is a particularly important change in the frequency of the MJO over the Indian Ocean (phases 2 and 3) and Maritime Continent (phases 4 and 5) between month 3 and month 4. Since the MJO teleconnections are strongly dependent on the phase of the MJO, these changes in the frequency and location of the MJO should impact the extratropical weather statistics. In addition, the increased MJO activity over the western Pacific is likely to increase the frequency of westerly wind bursts which can trigger oceanic Kelvin waves impacting the occurrence and amplitude of El Niño events. By contrast, S4 already overestimates the Maritime continent and W Pacific phases in the first month, while underestimating the Indian Ocean phase.

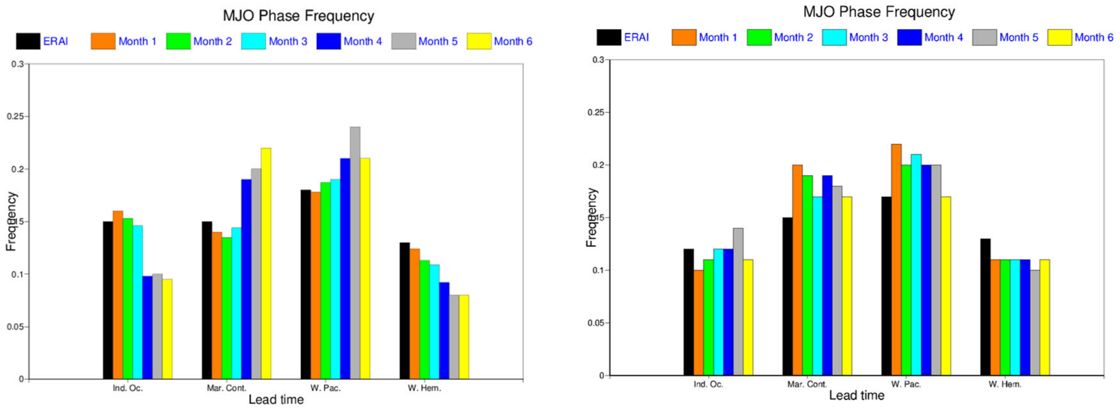


Figure 21: Ratio of days with a strong MJO (amplitude larger than 1) over the Indian Ocean (Phase 2-3), Maritime Continent (Phase 4-5), western Pacific (Phase 6-7) or western Hemisphere (Phase 8-1) in January, February and March in ERA Interim (black bars) and for various forecast lead times: month 1 (orange), month 2 (green), month 3 (cyan) and month 4 (blue). SEAS5 (left) and S4 (right).

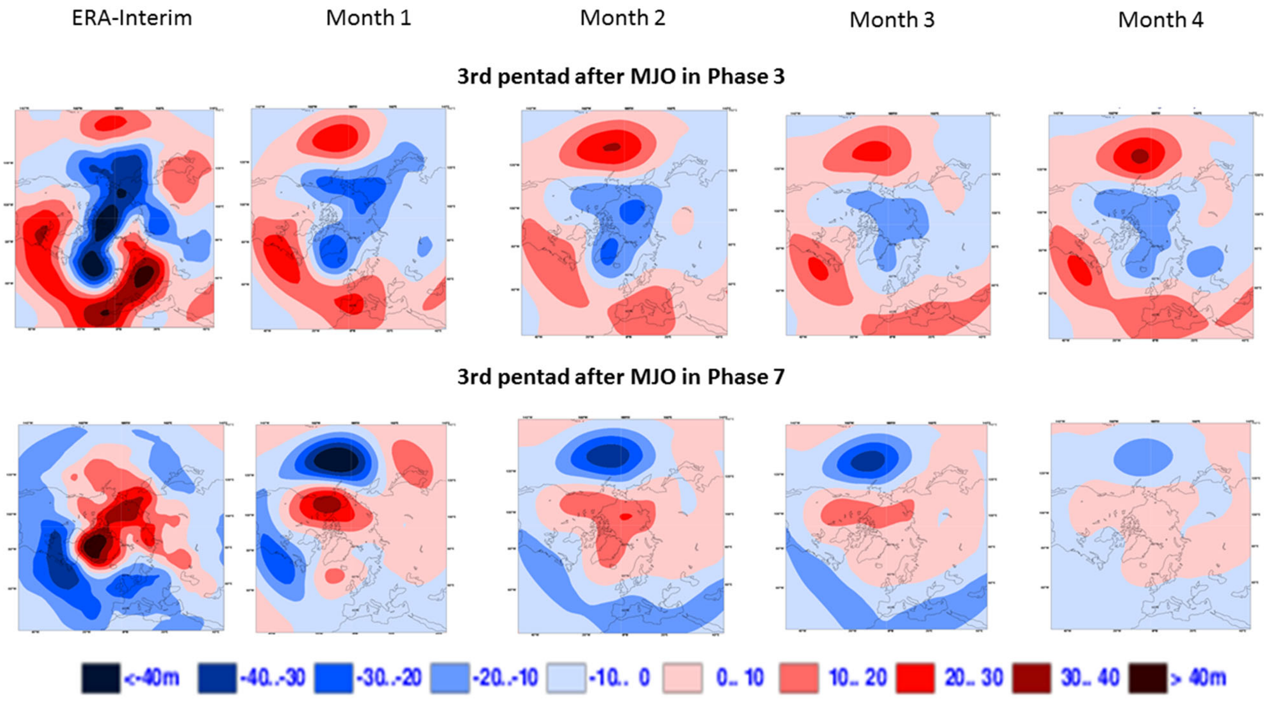


Figure 22: Composites of 500 hPa geopotential height anomalies (relative to model climate) 2 pentads after an MJO in Phase 3 (top panels) or Phase 7 (bottom panels) in ERA Interim (left panels) and in System 5 at various lead times (from month 1 to 4).

The impact of the MJO in the extra-tropics has been assessed by compositing the geopotential height anomalies at 500 hPa the third pentad after an MJO in Phase 3 (active phase of the MJO over the Indian Ocean). The delay is due to the time it takes for the Rossby wave generated by the MJO to reach the Euro Atlantic sector and impact the NAO (Cassou, 2008). According to *Figure 22*, the structure of MJO teleconnections are fairly well preserved during 6 months of integrations, but the teleconnections become increasingly weak over the Euro Atlantic sector as lead time increases. Therefore, the impact of the MJO on the NAO is likely to be significantly underestimated by SEAS5. Over the North Pacific, the MJO teleconnections are too strong, although this error tends to decrease with lead time after an MJO in Phase 7.

3.2.6 Exploring the impact of resolution and stochastic physics

We now discuss dependencies on resolution of atmosphere, ocean model and ocean initial conditions. *Table 4* lists the relevant experiments, some of which have been referred to previously. Our primary interest in investigating resolution has been to increase our understanding of SEAS5 and inform future development. An additional interest was in developing an affordable low-resolution configuration of the model for general research and development testing. A proper discussion of this is beyond the scope of this paper, but a TCo199 atmosphere resolution combined with an ORCA1_Z75 ocean is now our preferred general purpose low-resolution configuration.

Table 4: Different resolution forecast experiments based on SEAS5

Name	**Description**
SEAS5	High res. atmos (TCo319), high res. ocean (ORCA025_75)
SEAS5-lr **(low resolution)**	Low res. atmos (TCo199), low res. ocean (ORCA1_Z42) Ocean initial conditions produced by system similar to ORAS5 but at low resolution.
SEAS5-mr **(mixed resolution)**	High res. atmosphere, low resolution ocean, low resolution ocean initial conditions
SEAS5.ORCA1_Z75 **(SEAS5-or)**	High res atmos, low res 75 level ocean (ORCA1_Z75). Interpolated ORAS5 ocean initial conditions - see Appendix A.16 for details
SEAS5.ORCA1_Z42	High res atmos, low res ocean (ORCA1_Z42). Interpolated ORAS5 ocean initial conditions.
SEAS5-NoSP	SEAS5 without stochastic physics
SEAS5-lr-NoSP	As SEAS5-lr without stochastic physics
S4-NoSP	As S4 without stochastic physics
SEAS5-NoOobs **(CRTL-SST)**	As SEAS5, but ocean initial conditions produced without assimilation of in-situ ocean observations, altimeter and sea-ice. Only SST and surface fluxes are used.
CRTL-NoSST	As SEAS5, but ocean initial conditions from an ocean simulation forced by atmospheric fluxes – no data assimilation, no SST relaxation.

Impact of resolution and stochastic physics in the tropics

Resolution experiments are listed in *Table 4*. When ocean resolution is changed, we make an important distinction about the source of the ocean initial conditions – either a separate ocean analysis a low resolution, or a "dynamical interpolation" of the original ORAS5 analysis. Appendix A.16 contains details on this. Forecast experiments are for May and November starts over the 1981-2010 period, typically with 25-member ensemble.

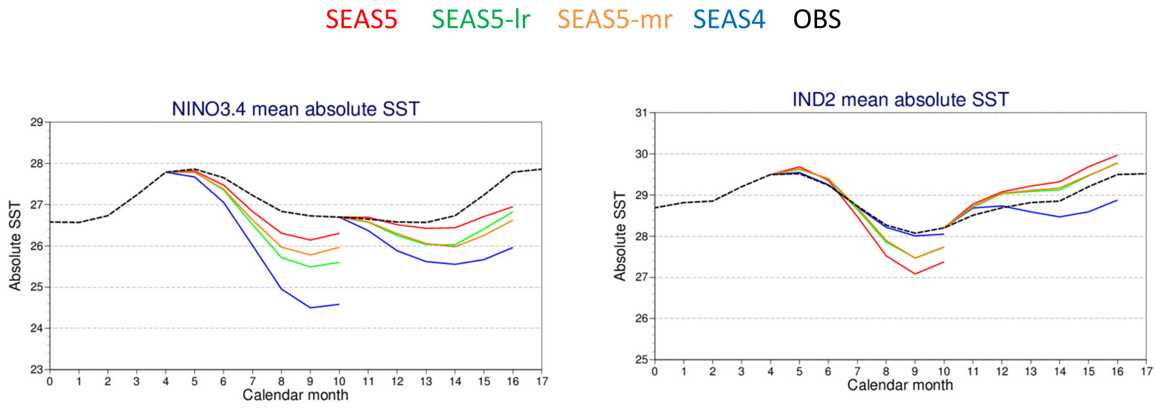


Figure 23: Impact of ocean and atmospheric resolution in SST bias in Nino3.4 and IND2

Comparing experiments SEAS5-lr and SEAS4 in *Figure 23* we estimate that about half the improvement in NINO3.4 SST prediction comes from the improved model. The other half is contributed by the increased resolution of SEAS5, which further reduces the SST biases in NINO3.4. This is accompanied by better interannual variability and forecast skill (not shown). The improvement coming from the increase in ocean resolution is related to a more realistic thermocline feedback (defined as the sensitivity of NINO3.4 to zonal wind stress perturbations). The resolution of the atmosphere also contributes, especially for forecasts initialized in May, which appears to be a consequence of a reduction of the zonal wind bias in the western Pacific. Stochastic physics also reduces the NINO3.4 biases, as discussed in Appendix A.15.

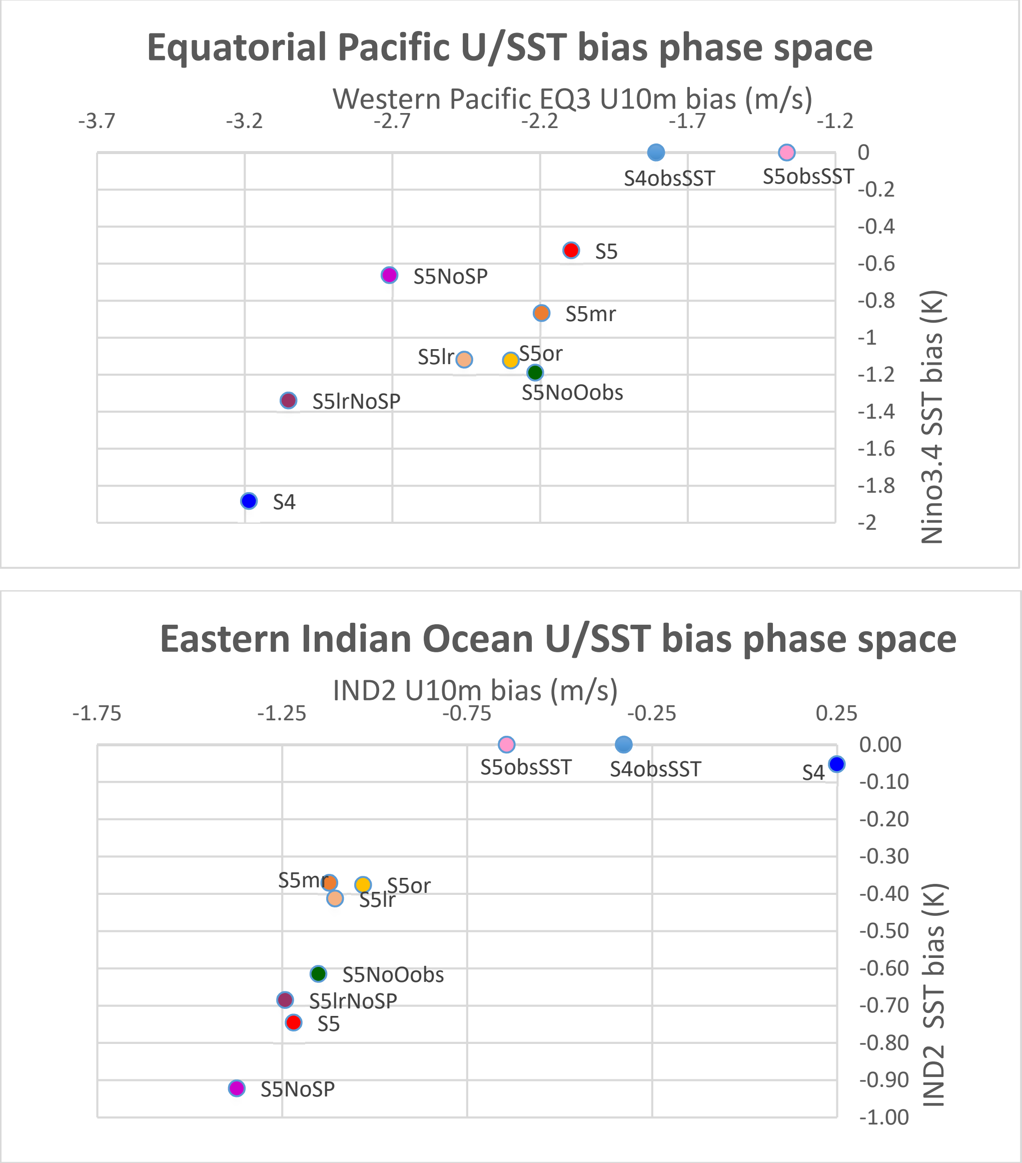


Figure 24: Relation between the biases in U10m and SST biases for the sensitivity experiments addressing the impact of resolution, ocean initial conditions and stochastic physics. The values correspond to forecast initialized in May and verifying in August (lead-time 4 months). Top) Remote EQ3-U10m and NINO3.4 SST bias. Bottom) Local U/SST bias in IND2. S4 is also shown. It can be seen that the coupled dynamics is very different in S4 from SEAS5.

Figure 24 (top) illustrates the relation between U10m bias in the Western Pacific (region EQ3) and SST biases in NINO3.4 in different experiments described in Table 6. Changing the ocean resolution mainly impacts NINO3.4-SST, with a slight impact on EQ3-U10m via a coupled Bjerknes feedback. The atmospheric resolution has a larger impact directly on EQ3-U10m, and indirectly in NINO3.4-SST. Changing the ocean initial conditions mostly changes the SST response (see displacement along the y-axis for experiment SEAS5-or with respect to SEAS5-mr and experiment SEAS5-NoOobs with respect to SEAS5). This impact of ocean and atmospheric resolution is in stark contrast with the impact of stochastic parametrizations (experiments SEAS5-NoSP and SEAS5-lr-NoSP). Deactivating the SP increases the EQ3-U10m bias substantially, with an indirect effect on NINO3.4-SST biases.

We previously noted that SEAS5 SST forecasts for region IND2 are worse than S4 (*Figure 4*b and *Figure 6*), especially for forecasts initialized in May: SEAS5 has a much stronger cold bias, an overestimation of the interannual variability, large ensemble spread, and poor forecast performance. The cold bias in SEAS5 is consistent with the easterly wind bias during this season. The relation between local U10m winds and SST in IND2 for the different experiments is illustrated in *Figure 24*b. Increased ocean resolution results in stronger upwelling and enhances the cooling, making the bias worse (note cluster of points in *Figure 24*b corresponding to the experiments with low ocean resolution, clearly separated from SEAS5 along the vertical axis). S4 is very different and had a westerly wind bias. The SEAS5 wind bias is relatively insensitive to atmospheric resolution, but does seems to depend on the atmospheric physics. Without stochastic parametrization the SEAS5 error would be substantially larger.

In terms of SST forecast skill, the performance of the low-resolution model (not shown) is in most places similar to SEAS5, except for NINO3.4 (worse with low resolution ocean), IND2 (better with low resolution ocean) and (outside of the tropics) the North-West Atlantic (better with low resolution ocean). These are also the main areas where the SEAS5 SST forecast skill differs from S4. In the central Pacific, the increased cold tongue associated with low ocean resolution leads to overestimation of the interannual variability, affecting both the amplitude of the ensemble mean and ensemble spread. By contrast, removing stochastic physics does not have a large impact on the amplitude of the interannual variability, but increases the RMSE and reduces the ensemble spread (see Appendix A.15).

Impact of resolution in the mid-latitudes

Changing the atmospheric model resolution to TCo199 has a minimal impact on tropospheric model climate, although does lead to a reduction in tropical lower stratosphere temperature biases, and improved Indian Ocean to mid-latitudes teconnections. It is rather the ocean resolution which has the bigger impact as part of the resolution upgrade between S4 and SEAS5 (see Appendix A.16).

Experiments have separated the impact of the higher resolution ocean model *per se*, and the high-resolution ocean analysis used to provide initial conditions to SEAS5. Details of these are given in Appendix A.16. Two important conclusions are that the horizontal resolution of the ocean is important for SST biases, but the vertical resolution makes very little difference, at least with the present model version; and that the higher resolution allows a reduction in SST bias in regions such as the North Atlantic, and that this has some impact on other atmospheric fields, although no impact on Atlantic blocking was seen.

It is initially surprizing that despite the improvements to model climate associated with increased ocean resolution, the skill in the northwest Atlantic, is systematically lower in SEAS5 than in S4. To shed light on this, we use our various SEAS5 resolution experiments to investigate. *Figure 25* shows the skill in the North-West Atlantic for SEAS5, S4, and the different variants of SEAS5 low and mixed resolution experiments. The lines clearly cluster in two distinctive groups, those initialized from the low-resolution ocean reanalyses, showing skill comparable with S4, and those initialized by ORAS5, with skill similar to SEAS5, irrespective of the horizontal or vertical resolution of the forecast ocean model. These results clearly demonstrate that the SEAS5 skill degradation over the North-West Atlantic is not due to the ocean model resolution during the forecast, but instead originates from the ORAS5 ocean initial conditions. This is sufficiently important that we consider it in detail in the next section.

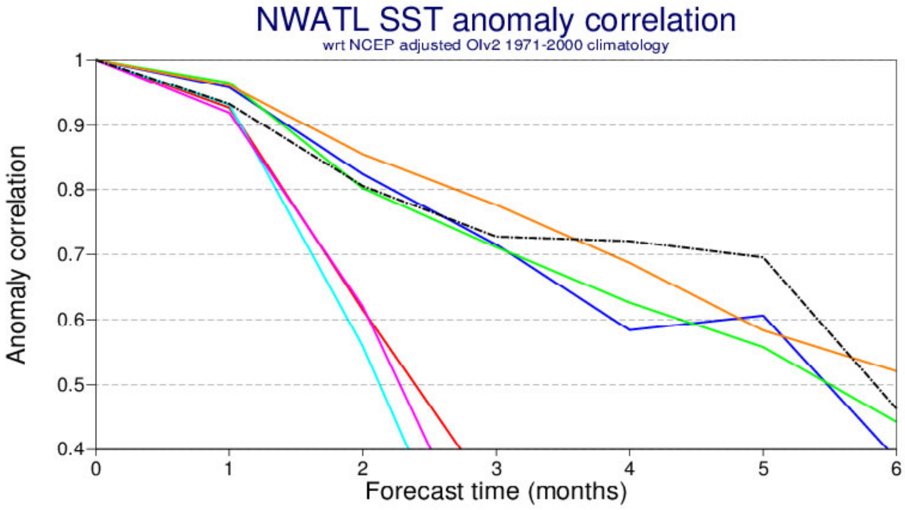


Figure 25: Anomaly correlation skill for forecasts of SST averaged over the NW Atlantic, for November forecasts in the period 1981-2015. SEAS5 (red), S4 (blue), SEAS5-lr (green), SEAS5-mr (orange), SEAS5.ORCA1_Z42 (cyan), SEAS5.ORCA1_Z75 (magenta). Forecasts based on ORAS5 initial conditions have poor skill, regardless of resolution.

3.2.7 The North Atlantic problem

There is a pronounced skill deterioration for winter (DJF) surface temperature in parts of the North Atlantic in SEAS5. The affected region is centred on a box defined by the longitudes 50-30W and the latitudes 45-55N, which we henceforth refer to as the North-Atlantic skill deterioration (NASD) region. This region is characterized by complex interactions of several large-scale ocean currents that are key to the North Atlantic Ocean circulation (Buckley and Marshall, 2016). The deterioration of skill in this region can potentially affect forecasts over Europe through advection by the prevailing westerly winds. It is therefore important to understand this problem, to establish its impact on atmospheric variables, and find a remedy before the implementation of the next seasonal forecasting system.

The principal contribution to the skill degradation comes from a non-stationary bias of SEAS5, rather than an inability to forecast interannual variability. *Figure 26* shows the spatial pattern of the SST bias for S4 and SEAS5 for the early period 1981 - 1995 and the late period 2001-2015. A constant positive SST bias in the Gulf Stream region is present in both S4 and SEAS5. This is connected to a well-known failure of low-resolution ocean models to simulate the separation of the Gulf Stream from the North American coast correctly (Chassignet and Marshall, 2008). Note that this bias has improved throughout in SEAS5 thanks to the increased ocean resolution. The problem occurs further downstream in the NASD region, where the Gulf Stream meets with the cold Newfoundland Current coming from the North, and splits into the North-Atlantic Subtropical Gyre and the North Atlantic Drift. In this region

S4 had a persistent cold bias in this region thorough the record, whereas SEAS5 has a strong warm bias in the early period and a small negative bias in the later period. While the constant bias in S4 is essentially removed in the bias-corrected forecast products, the non-stationary bias of SEAS5 is not. The forecasts from SEAS5-lr and SEAS5-mr, which use the low-resolution ocean initialized from a low-resolution ocean reanalysis, show stationary bias like S4. In contrast, the non-stationary bias seen in SEAS5 appears in forecasts using the low-resolution ocean initialized from ORAS5 (SEAS5.ORCA1). The non-stationarity of the bias is the key factor in the skill difference shown in *Figure 25*.

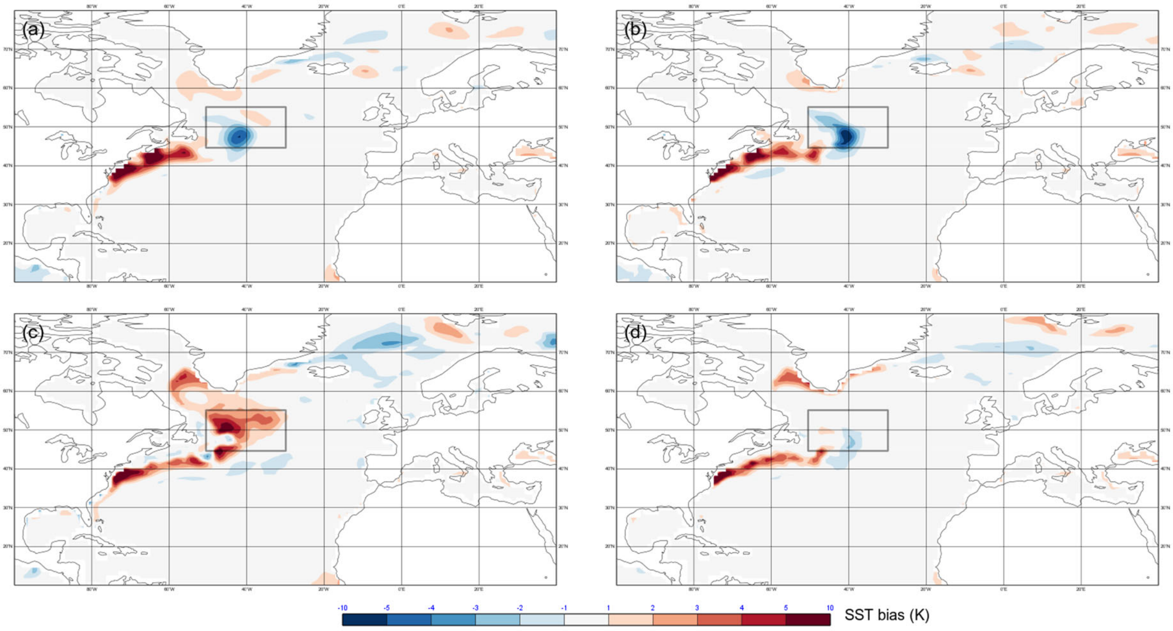


Figure 26: DJF SST bias for November forecasts w.r.t. ERA-Interim for (a), (b) S4 and (c), (d) SEAS5. The bias during the early period 1981-1995 is shown in (a, c), the bias during the late period 2001-2015 is shown in (b, d).

The non-stationary bias in SEAS5 comes from the ocean reanalysis ORAS5 that provides the ocean initial conditions for the reforecasts. As with previous ECMWF ocean reanalyses, ORAS5 is constrained to observed SST by imposing a damping heat flux of 200 W m^{-2} K^{-1}. In the early period before the mid-1990s, in-situ observations of ocean temperature and salinity are too sparse to constrain the ocean state efficiently. ORAS5 then exhibits strong sensitivity to the formulation of the SST constraint, demonstrated by running two experiments where the forecast model is identical to SEAS5, but where the initial conditions are taken from ORAS5 control simulations with the data assimilation switched off. The only difference between the two control simulations is whether SST relaxation has been activated (Ctrl-SST) or deactivated (Ctrl-noSST). *Figure 27* clearly shows that the warm bias in the early period for reforecasts started from Ctrl-SST is even worse than in SEAS5, whereas Ctrl-noSST is virtually bias-free during the early period. In the late period, the bias in Ctrl-SST becomes smaller, whereas Ctrl-noSST develops a strong cold bias.

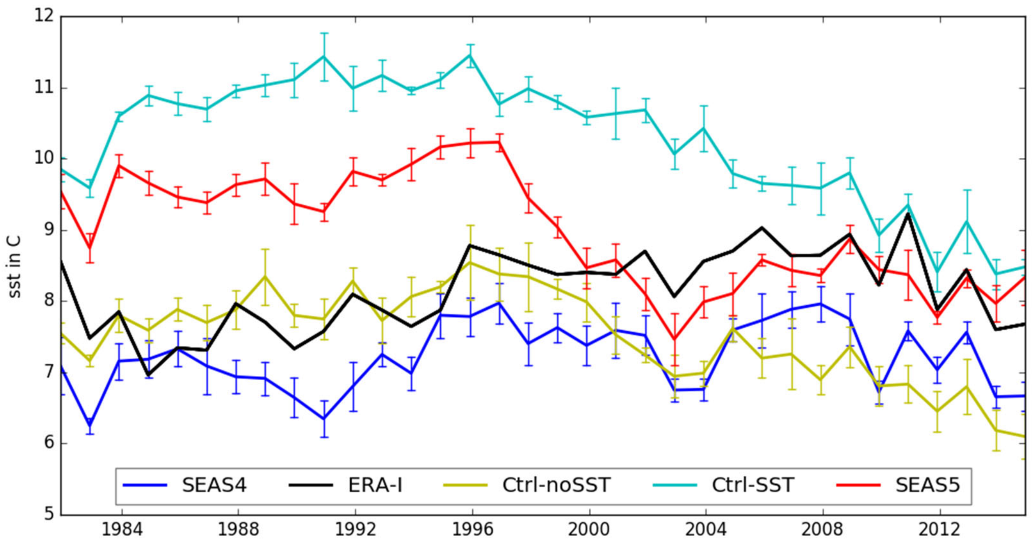


Figure 27: Time series of DJF sea-surface temperature over the NASD region 50-30W, 45-55N. In black are observations represented by ERA-Interim, and coloured lines represent various reforecast sets (solid line connects ensemble means, and error bars denote ensemble spread). The operational forecasting systems shown are S4 in blue and SEAS5 in red. Shown in yellow (Ctrl-noSST) and cyan (Ctrl-SST) are experimental reforecasts, where the forecast model is identical to SEAS5, but the ocean initial conditions are taken from an ocean simulation without data assimilation, where the relaxation to observed SST is either activated (Ctrl-SST) or deactivated (Ctrl-noSST).

The SST variability in the North Atlantic is closely linked with the strength of the meridional transports of heat and fresh water. A good indicator for the strength of this meridional ocean transport is the Atlantic Meridional Overturning Circulation (AMOC) index, which measures the volume transport of ocean water in the upper 1000m of the North Atlantic at 26N. *Figure 28* shows that there is very high (r = 0.88) correlation between the DJF forecast SST in the NASD region and the analysed AMOC strength in the year the forecast was started. This remarkable result demonstrates that the representation of the AMOC, traditionally considered a mode of decadal variability, can have a direct impact on seasonal forecasts even within the first few months. Our different ocean analyses vary in their representation of AMOC variability. Although we would like to confirm which analysis is most accurate, it is difficult to ascertain the realism of the AMOC strength for the full reforecast period – see Appendix A.17.

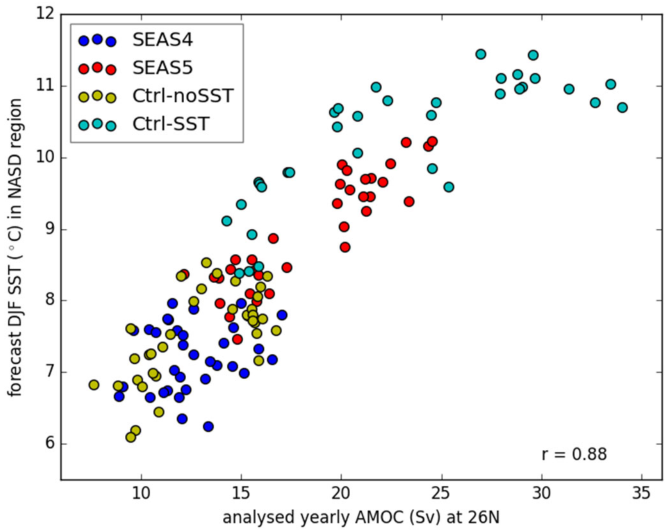


Figure 28: Scatter plot of DJF forecast SST in the NASD region and the annual-mean AMOC in the analysis that each forecast is started from. The Pearson product correlation across all forecasts is 0.88.

The mechanism linking the SST relaxation to enhanced ocean transports and warm SST bias in the NASD region is still under investigation. However, there is a large body of literature linking enhanced buoyancy loss in the NASD region to increased overturning circulation (see Buckley and Marshall, 2016). Figure 29 demonstrates that the additional heat flux required to keep ORAS5 close to observed SST corresponds to an extremely strong (in excess of 600 Wm^{-2}) additional cooling of the region in question, which corroborates the hypothesis of an established positive feedback loop of compensating errors in the early period of ORAS5, where the region of the NASD region receives a surplus of heat from an unrealistically strong Gulf Stream, which triggers strong additional buoyancy loss from the relaxation to observed SST, which in turn invigorates the Gulf Stream.

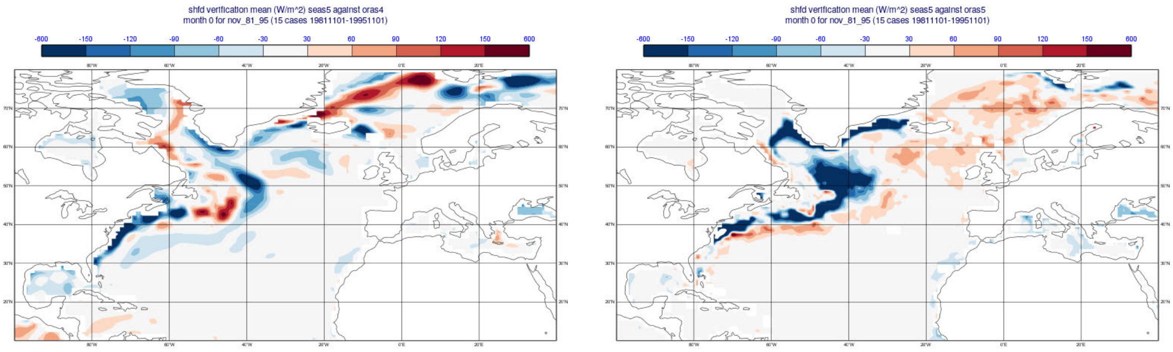

shfd verification mean (W/m^2) seas5 against oras4
month 0 for nov_81_95 (15 cases 19811101-19951101)
shfd verification mean (W/m^2) seas5 against oras5
month 0 for nov_81_95 (15 cases 19811101-19951101)
-600
-150
-120
-90
-60
-30
30
60
90
120
150
600

Figure 29 Additional surface heat flux instigated by the relaxation to observed SST for November during the early period 1981-1995 for (left) ORAS4 and (right) ORAS5.

It is difficult to infer the root cause of this error, but the balance between excessive ocean transports and extremely strong SST relaxation keeps ORAS5 reasonably close to the SST available observations, despite the implied biases in water mass properties and ocean transports. However, the time scales of the two processes are very different: the SST relaxation is switched off immediately when the SEAS5 forecasts start, while the time scale of adjustments in ocean transport is much longer than the forecast lead time. Therefore, in the SEAS5 and Ctrl-noSST forecasts the excessive ocean transports still provide excess warming to the NASD region, while the cooling from the SST relaxation is absent, leading to the problematic warm bias in the SST. This line of reasoning is supported by the spatial pattern of the SST relaxation in ORAS5 shown in Figure 29: it correlates extremely well with the pattern of DJF SST bias in SEAS5 reforecasts.

Wider impact of North Atlantic SST errors

The impact of North Atlantic SSTs on the atmospheric circulation is complex, and long time series are required to obtain statistically robust results (e.g. Czaja and Frankignoul, 2002). Observation-based and modelling studies suggest that North Atlantic Subpolar Gyre SST anomalies play a role in forcing the atmosphere (e.g., Robson et al, 2013 and reference therein). Indeed, the prediction of the decadal variability of this region is a corner stone of decadal prediction activities. Less attention has been paid to the role of this region for seasonal forecasts.

To attempt to isolate the impact of the NASD SST errors, we compare period differences from SEAS5 and from its low-resolution analogue SEAS5-lr that has a stationary SST bias in the North Atlantic. The early period is defined as 1981-1995 and the late period as 2001-2014. If $S5_l$ and $S5_e$ are the biases of SEAS5 in the late and early periods, respectively, and LR_l, LR_e the corresponding biases in the low-resolution analogue, the difference Δ is written as $\Delta = (S5_e - LR_e) - (S5_l - LR_l)$. The period difference removes the mean differences between SEAS5 and SEAS5-lr, leaving only the differences in decadal variability or trends. As expected, computing Δ for SST shows a dominant positive signal over the NASD region (not shown).

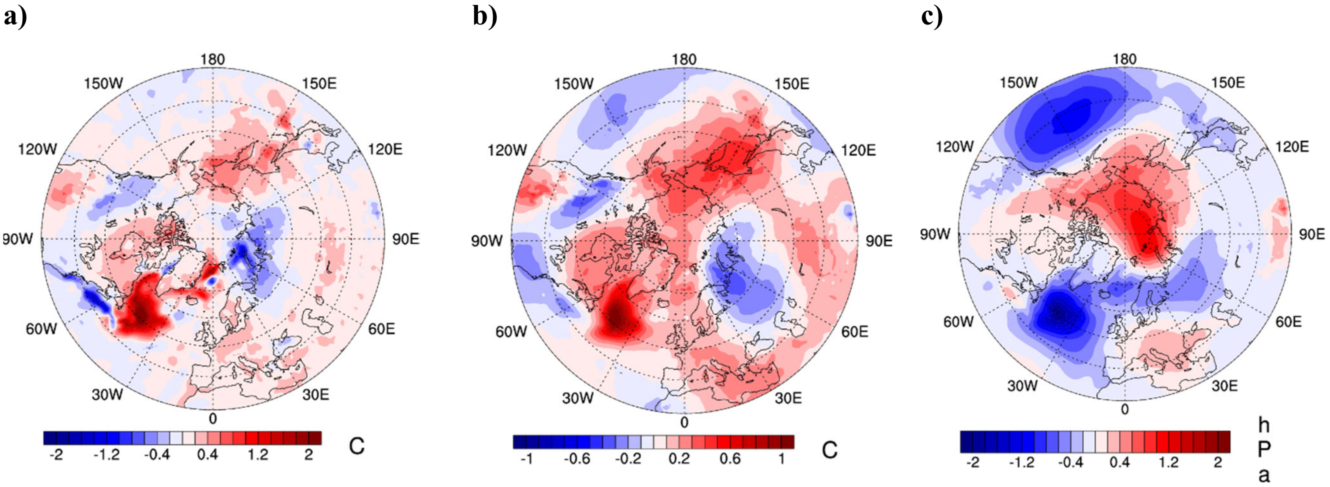


Figure 30: Period difference between 1981-1995 and 2001-2014 averages of differences between SEAS5 and SEAS5-lr DJF bias in a) 2m temperature, b) 850hPa temperature, and c) mean sea level pressure

Results are shown for different atmospheric parameters in *Figure 30*. Period differences Δ for 2m temperature exhibit a clear maximum of around 2K over the NASD region, with more positive differences in the early period, indicating the direct impact of the SSTs via local air-sea fluxes. Period differences in 850hPa temperature show a similar albeit weaker maximum in the NASD region, which seems to spread out further downstream over Europe, although the amplitude of the signal is small compared to the level of interannual variability. Period differences in mean sea level pressure (MSLP) show a pronounced minimum in the NASD region. This is consistent with the findings by Booth et al (2012), who report more intense cyclogenesis in the North Atlantic associated with warmer SSTs.

The suggestion that the NASD SST evolution has a broader impact on the atmosphere in the North Atlantic region is supported by a CCA analysis of the co-variability of SEAS5 SST and precipitation errors – see Appendix A.18 for details. It is also supported by the fact that SEAS5-lr, which does not exhibit the NASD problem, has somewhat higher DJF skill than SEAS5 in this region (Figure A 17 in Appendix A.16). Dedicated numerical experimentation would be needed to investigate further.

An important question is how to ameliorate or solve the problem. The results show a strong relationship between the temporal evolution of the SEAS5 bias and the AMOC in the ocean initial conditions, which in turn is dependent on the SST relaxation. The strong relaxation coefficient has been used in all previous ECMWF ocean reanalyses without trouble, but the sensitivity to the SST relaxation appears stronger in the ORCA025 configuration, perhaps because at higher resolution the model is able to reproduce a stronger boundary current. The AMOC also appears sensitive to the model bias correction used during data assimilation, and is probably sensitive to the parameterization of deep convection in the NEMO model. These aspects are being looked at in preparation for the next ocean reanalysis. Improved methods for assimilating SST information are also needed, which we are pursuing in collaboration with NEMOVAR partners.

3.2.8 QBO and stratospheric teleconnections

The quasi-biennial oscillation of the tropical stratosphere (QBO, Reed et al, 1961) provides one of the few purely atmospheric sources of predictability on the seasonal timescale, and is a phenomenon which S4 could predict relatively well. Initial experiments with Cy43R1 showed that this would not be the case if the default settings for the parametrization of non-orographic gravity wave drag (NOGWD) were used, and this led to a reduced tropical value being used, as described in Section 2.1.

To illustrate the motivation for and result of this change, we compare the amplitude and phase of the QBO as a function of lead time for S4, the default Cycle 43r1 IFS, and SEAS5 in *Figure 31*. We use the monthly zonal wind from 5°N to 5°S as a QBO index, looking first at the 30 hPa level. The anomaly correlation of the default IFS Cycle 43r1 shows a large decrease after month two for both May and November starts, and is in fact no better than a persistence forecast. This is because the phase propagation of the QBO is much too fast. By reducing the amplitude of the tropical NOGWD the QBO phase propagation is slowed. Although it was possible to achieve a reasonable phase propagation of the mid-level QBO in SEAS5, this does not mean that the QBO is well-modelled as a whole. This is notable in the amplitude of the QBO, which at 30 hPa is damped even more strongly than S4, with typically only half the observed amplitude after 7 months. The NOGWD has been tuned to preserve a comparable level of skill in predicting phase (thus avoiding the risk of damaging false signals), but the amplitude of the QBO in the lower stratosphere, which was already weak, is even further reduced.

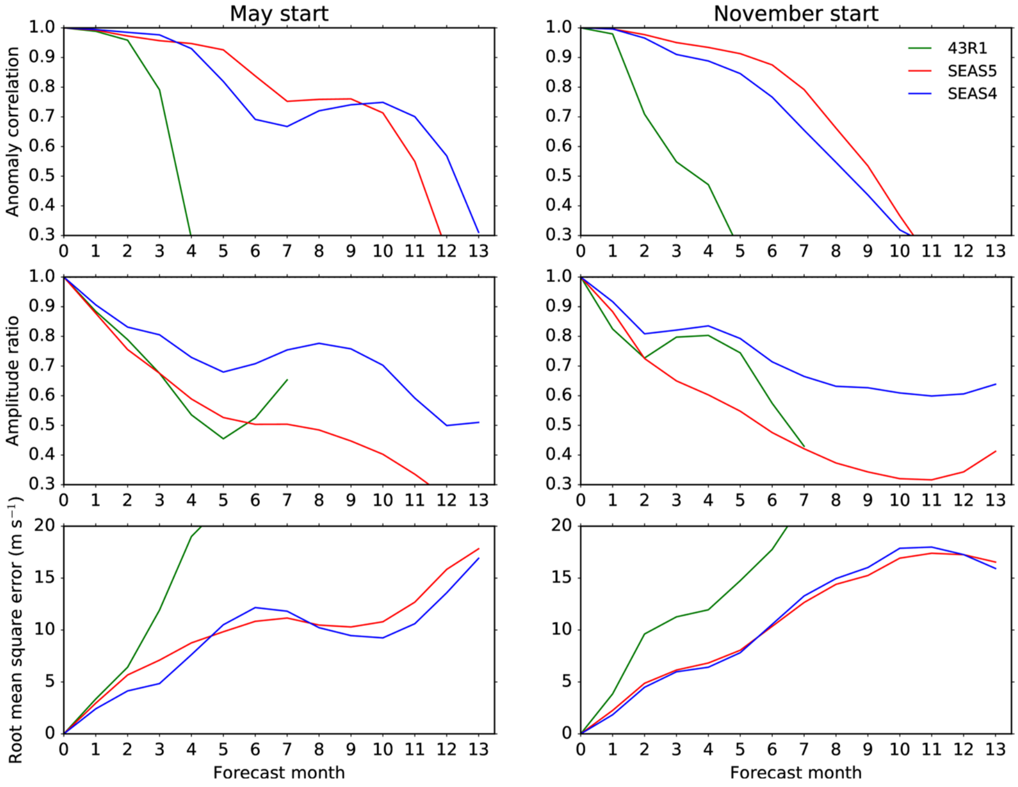


Figure 31: Metrics summarising the phase and amplitude of the QBO at 30 hPa in S4 (blue), SEAS5 (red) and IFS cycle 43r1 with default settings (green) relative to ERA-Interim reanalysis for forecasts initialised in May and November, using five ensemble members with initialisation dates from 1981 to 2016. Top panels: Anomaly correlation, middle panels: ratio of the standard deviation of the system to the standard deviation of ERA-Interim reanalysis, bottom panels: RMSE.

We can also look at teleconnections from the QBO. Observations suggest a fairly strong relationship with the NH winter polar vortex (Holton-Tan effect), and also a connection to the NH surface circulation. SEAS5 reproduces Holton-Tan effect very weakly, less well than S4. The connection to the surface is also reproduced, but is again too weak. An L137 version of SEAS5 improves the teleconnections significantly, but they remain too weak. A summary is given in Appendix A.19.

Stratosphere-troposphere coupling and the polar vortex

Recent work has shown that some seasonal forecasting systems such as S4 suffer from anomalously low signal-to-noise ratios in forecast of NH winter circulation indices such as NAO and Northern Annular Mode (NAM). For example, Scaife et al (2014) and Stockdale et al (2015), discuss that ensemble mean hindcasts of large-scale extra-tropical circulation variables are well correlated with observations but with very low mean amplitude in comparison to the spread of the ensemble.

To try to shed more light on the origin of the low signal-to-noise ratio, we performed an analysis of the SEAS5 re-forecasts using a newly published Bayesian statistical model for seasonal predictability (Siegert et al, 2016). In common with previous models, SEAS5 shows anomalously low signal-to-noise ratio for the NAM (one measure of the large-scale extra-tropical climate state) when initialised in November. However, this low signal-to-noise ratio is not present for forecasts initialised in December and is only present from the lower stratosphere to the surface (see Figure 32). Further, forecasts of the winter mean meridional heatflux at 100 hPa (a proxy for the upward propagating Rossby wave flux) do not show anomalously low signal-to-noise ratio for forecasts initialised in November. Taken together, these pieces of evidence suggest that the extra-tropical, lower stratospheric bias which develops in the model over the first two weeks of integration may play an important role in limiting the size of the predictable signal that can be produced by the model in the troposphere.

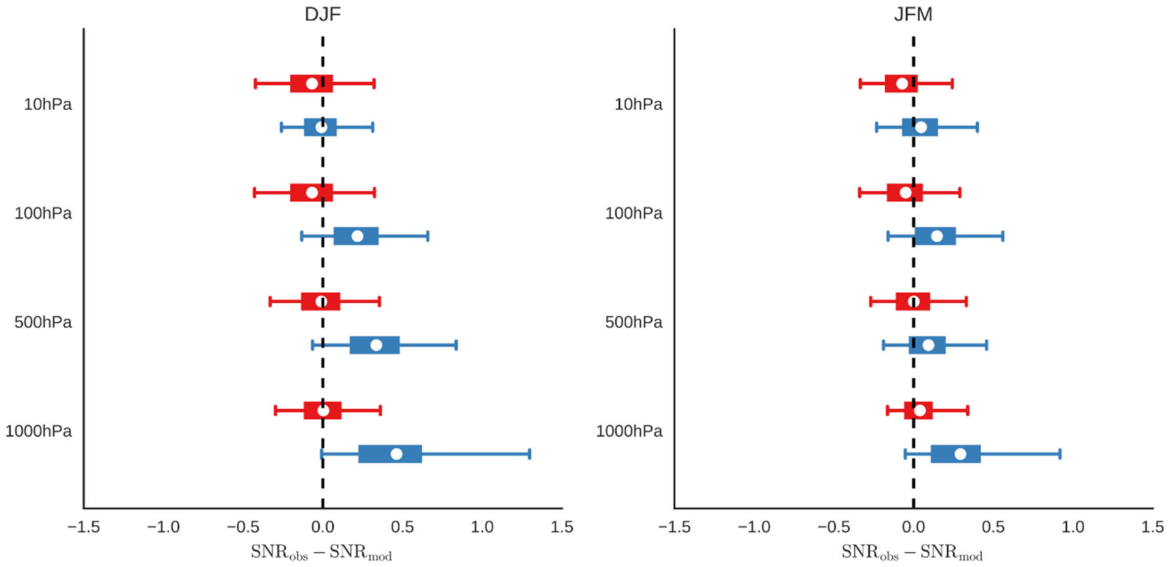


Figure 32: Box plots showing posterior estimates of the difference between the signal-to-noise ratio of the observations and model (SNRobs - SNRmod) as a function of model level and season. The hindcast set used is from 1982-2016 for SEAS5 with 25 ensemble members. In the box plots, the central box shows the (25,75) credible range of the difference between signal-to-noise ratio, the white dots show the mean and the whiskers show the (2.5,97.5) credible range. Blue shows the 1 November starts, red the 1 December starts.

3.2.9 Predicting the onset of large warm ENSO events

Of particular interest for seasonal forecasting is the prediction of the onset of ENSO (typically during summer) from forecasts initialized in spring. This is a challenging prediction, since the forecast crosses the so-called “boreal spring predictability barrier”, when the predictive skill of dynamical and statistical models significantly drops, regardless of the initial conditions. Reasons for the predictability barrier are various. See for instance Duan and Hu (2016), and references therein. The coupling between surface and subsurface ocean is weaker at this time of the year, the SST gradients along the equator are reduced, and the coupled system is more unstable and susceptible to weather noise and westerly wind bursts (WWB). It is also the time of the year when the interannual variability of SST is at its minimum. The organized convection over the Western Pacific is influenced by both sub-seasonal variability (MJO is still active) and the seasonal cycle (the ITZC crosses the Equatorial Western Pacific). A key element for seasonal forecasts at this time of the year is predicting the location and strength of organized deep convection. If the deep convection remains at the Equator and is displaced westward, then it is very likely that a warm event develops.

In the past, the failure of the ECMWF to represent the MJO and associated WWB led to failures in the prediction of the onset of ENSO (Vitart et al, 2003). This can be seen in the left panel of *Figure 33*, which shows the prediction for the 1997/98 El Niño event by successive ECMWF forecasting systems. It was not until S4 that the coupled model was able to generate WWB associated with the MJO, but because of the errors in the S4 mean state, the interannual variability was overestimated. The prediction of this event by SEAS5 improves over S4, with more realistic amplitude. The same successful story can be told for the prediction of the recent 2015/16 El Niño (lower left panel in *Figure 33*).

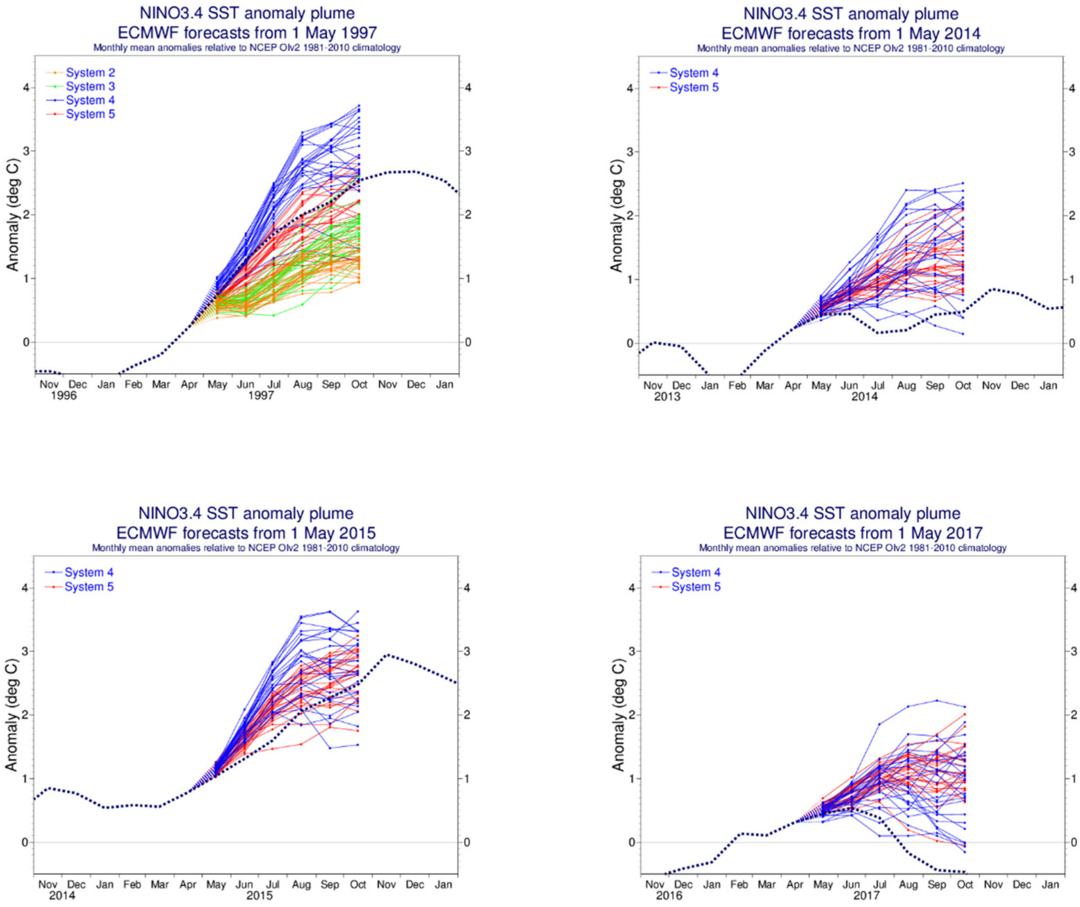


Figure 33: Forecast plumes for the largest forecast anomalies in SEAS5 from forecast initialized in May, namely 1997, 2014,2015 and 2017. Shown are the anomalies for SEAS5 and S4, and for May 1997 the forecast from SEAS2 and SEAS3 are also shown.

The overestimation of variability in S4 led to large ensemble spread and overprediction. It has been speculated that this was the reason for the apparent false alarms during 2014 and 2017, where S4 predicted the chance of warm events that did not occur –although the ensemble spread was large. SEAS5 has relatively good variability during this season. Still, it also produced forecast similar to S4 for 2014 and 2017, albeit with reduced spread. Whether these forecasts were false alarms or instead the system was quite unpredictable still needs to be assessed. Visual inspection indicates that the forecasts for 2014 and 2017, when a large warm event did not occur, are different from those of 2015, when a warming actually happened. During the reforecast period of SEAS5 there have been only three large warm events in observations: 1982, 1997 and 2015. These were successfully captured by SEAS5. The limited number of occurrences prevents a proper probabilistic assessment. The recently completed seasonal reforecast for 110 years, which uses the SEAS5-lr initialized from CERA-20C should allow characterize better the reliability in the prediction of extreme large events.

Low frequency modulation of tropical SST errors

The fact that these perceived “false” alarms in both S4 and SEAS5 have occurred in the recent period, prompts the question on whether the errors in the seasonal forecasts are modulated by low frequency decadal variability or trends. *Figure 34* shows the temporal record of SST errors along the Equator for SEAS5 forecasts initialized in May, when verifying in May (left) and August (right). The figure shows the difference between normalized SST anomalies of model ensemble mean and observations. The normalization factor is the standard deviation of the interannual anomalies of the ensemble mean and observations respectively. In the recent years, SEAS5 consistently predicts warmer conditions than observations. This tendency is already visible in the first month into the forecast, and grows in time. Further analysis suggests that recent errors are associated with the failure of the model to reproduced observed trends in the circulation in the Eastern Pacific and Atlantic Ocean. According to ERA-Interim, there is a trend towards an asymmetric meridional mode, with warm SST anomalies north of the Equator and cold SST anomalies south of the Equator, and stronger cross equatorial northward winds (not shown). The model appears unable to capture the strengthening asymmetry, and produces instead symmetric warmings with equatorial maxima. This error in the trend might be related with the too-zonal behaviour of the mean atmospheric circulation discussed in section 3.2.3. Work is ongoing to farther characterize and understand low-frequency errors in the forecasts.

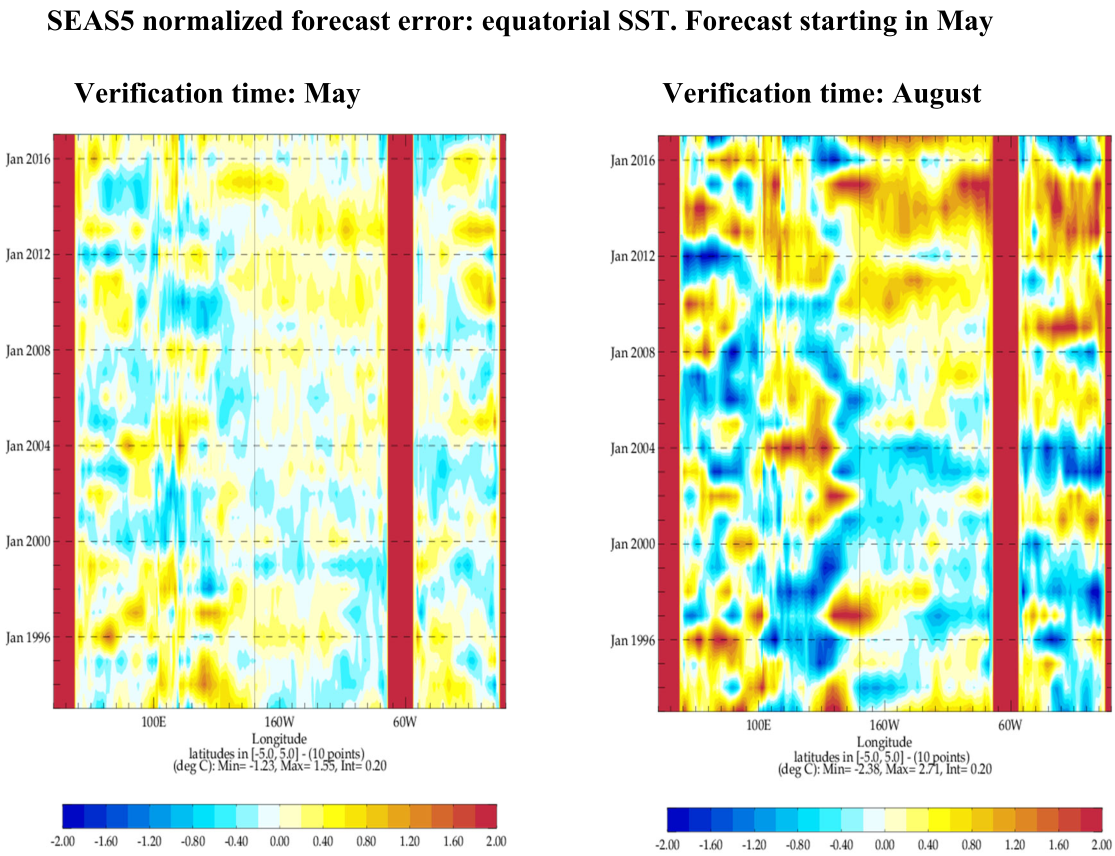


Figure 34: Longitude/time Equatorial section of differences of normalized SST anomalies between the ensemble mean of May-start SEAS5 forecast and analysis. Forecast verifying in May (left) and August (right) are shown. The forecasts are normalized by the interannual standard deviation of the ensemble mean for the corresponding lead times.

3.2.10 Remaining gaps in the evaluation

There is no such a thing as a complete evaluation of a forecasting system. The evaluation provided here focuses on aspects relevant for system development, and does not attempt to provide a user-based assessment. As such, it has given particular attention to the problems and weaknesses that have been identified. We note two important areas which have not been examined: an evaluation of land processes and comparison with other non-ECMWF forecasting systems.

The biases in T2m over land in seasonal forecasts typically show the same structure as those in the extended range forecast, allowing joint evaluation of changes in land processes in both seasonal and extended range systems. To understand the specific contribution of land initial conditions to seasonal predictability and errors, we plan a routine evaluation comparing the operational predictions from SEAS5 not only with those from SEAS5-ObsSST, but also those from an AMIP run equivalent to SEAS5-ObsSST, where the land initial conditions are unconstrained. This routine attribution of predictability and forecast sensitivity based on case studies should shed light into the role of the land initial conditions in our operational systems. ERA5 should provide a better reference against which to evaluate the model, but for surface fields it will not be perfect, and we will have a continuing need to seek reliable observationally-based land surface verification datasets.

In this evaluation of SEAS5 we have not included any comparison with other seasonal forecasting systems. Within the activity of Copernicus Climate Change Services an independent evaluation and comparison of forecast skill has been carried out using different seasonal forecast systems. However, the current evaluation considered S4, and at the time of writing a multi-system evaluation including SEAS5 is not available. Although the C3S evaluation activity is primarily user oriented and targeting multi-model products, we expect that the feedback received will improve our understanding of the strengths and weakness of SEAS5. Comparative multi-model studies are useful in that differences in performance highlight potential areas for improvement, although care will always be needed in interpreting scores due to the often large error bars from sampling. The public availability of seasonal data in the C3S climate data store should facilitate scientific studies, which will also help with the evaluation of SEAS5. If timely and well targeted, this community feedback could feedback on model development. Once the C3S Climate Data Store is fully up and running, this should also make it easier for us to produce our own routine comparisons with other systems.

Finally, we note that at the time of writing the re-forecasts for February, May, August and November start dates are being extended to 51 members. This will enable better sampled scores, and better comparisons with S4, for which a similar extension exists.

3.3 Summary of performance

We have presented the seasonal forecasting system SEAS5, which is a major upgrade in terms of model resolution and re-forecast data set, and includes prognostic sea ice for the first time.

Forecast scores show that progress in ENSO prediction is clear cut. Skill estimates based on the full set of start dates also suggests some progress in the northern extra-tropics, most clearly in 2m temperature forecasts and tropical scores, but it is harder to detect overall improvement in prediction of NH circulation anomalies, particularly for months 2-4 of the forecast. Assessment of incremental improvements in mid-latitude skill is generally more challenging, due to the low signal to noise ratio in the model (meaning that re-forecast ensemble sizes are still not adequate to determine the model signal) and the dependence of the scores on the verification period (the verification period is not long enough and/or stationary enough for us to be confident in estimating expected future skill from the past, e.g. Weisheimer et al, 2017).

The climate of SEAS5 has been characterized in terms of mean state, air-sea interaction in the tropics, stratosphere-troposphere interaction and evaluation of processes such as tropical cyclones, MJO and teleconnections. All these aspects have helped to obtain a broader picture of SEAS5 performance and challenges ahead for future developments, which we summarize below:

- **Noteworthy improvements to ENSO SST forecasts**, building on our previous world-leading skill. Both model developments and increased ocean resolution contribute to the improvements. The system is still not fully reliable, however, and longer re-forecast periods might be useful in further exploring the reliability of extreme warm events.

- **Dynamical sea-ice brings a new source of predictability in SEAS5**. The bias-corrected sea-ice forecasts show high skill in predicting sea-ice anomalies in the first few months of the forecasts, with positive effects in the prediction of T2m in surrounding areas. However, large biases develop especially in summer.
- **Differences in climate dynamics between SEAS5 and S4**:
    - Equatorial Pacific: Weaker cold tongue bias in Eastern Pacific and improved variability, thanks to atmospheric model improvements and higher horizontal resolution in the ocean. Still the cold tongue erodes the warm pool region, and not enough meridional asymmetry.
    - Indian Ocean: Eastern Indian Ocean shows cold/dry/easterly bias, reduced skill and large ensemble spread. These errors are enhanced by coupled feedbacks, and the high-resolution ocean.
    - Convection over the Maritime continent and MJO is improved, but appears skewed towards the Western Pacific. Whether this is related with the biases in the Eastern Indian Ocean, and the degraded Indian Ocean-mid latitude teleconnections need further investigation.
    - The North Atlantic shows enhanced errors in SST forecasts in the subpolar gyre, which appear to affect the behaviour of atmospheric fields over Europe. These errors are related to imbalances in the ocean initial conditions from ORAS5, introduced by the change in ocean resolution.
- **Some mean state improvements**, **but stubborn biases remain**: upgrades in the coupled model and increase in resolution result in many improvements in model climate, but some persistent errors remain:
    - The atmospheric circulation appears to be too zonal, which manifests in too strong equatorial easterlies. The eastern Indian ocean is of particular concern.
    - A marked zonally symmetric bias in Z500 centred at 40-45N in JJA.
    - Biases in the lower stratosphere and the sub-tropical jet may be limiting the correct propagation of teleconnections, as might the lack of QBO amplitude in the lower stratosphere.
    - The teleconnections from the Indian ocean over the North Atlantic sector have not improved.
- **Increased resolution can give rise to new issues**: increasing atmospheric horizontal resolution without sufficient vertical resolution has enhanced temperature errors in the lower stratosphere; the increase in the ocean resolution, although beneficial for many aspects, has also led to process imbalance, deteriorating the performance in the North Atlantic subpolar gyre and Eastern Indian Ocean.
- **Non-stationary forecast errors** are visible in the North Atlantic and Tropical Pacific.
    - This causes difficulties for bias corrected forecast products.

- The errors appear related to low frequency variability of the climate system, and indicate that skill gains can be made by improving the ocean initial conditions and model.

- **Positive prospects regarding mid-latitude seasonal predictability**: there is suggestive evidence that more realistic representation of tropical-midlatitude teleconnections, stratosphere-troposphere interaction and simulation of the QBO are possible. All of these are important predictability drivers at the seasonal time scales, which should lead to increase skill in the extratropics once upper-air biases are better controlled. Decadal variability, if well captured, can also contribute to seasonal predictability.

In the context of seamless forecasting systems, the evaluation of model cycles to assess seasonal forecast performance is a key building block. The climate evaluation protocol needs revising to better represent the model used in the forecasting system. For instance, it should include stochastic physics, and representative ocean initial conditions. It should also be updated to include ERA5 initialization. The metrics used for evaluating model climate also need further improvement. We plan to tackle these aspects in the next few years.

4 Looking to the future

4.1 Scientific priorities for developments of seasonal forecasting systems

There are many requirements for the continued development of our seasonal forecasts. Here we highlight a few priorities, while emphasizing that a broad range of model and assimilation improvements are needed to underpin future progress.

4.1.1 Weaknesses to be addressed

- The “North Atlantic” problem in the ORAS5 ocean re-analysis needs to be resolved, and a new ocean re-analysis carried out. Assimilation of SST may be a useful capability in this regard.

- Sea ice biases are large in both summer and autumn, and reducing these may allow further benefit from our sea-ice forecasting capability.

- Tropical wind biases are still problematic, and they have particular impact around the Maritime continent and west Pacific. Further improvements in the behaviour of convection and the MJO in this region are also considered important.

- Biases at the tropopause level and in the stratosphere should be reduced, and our understanding of the impact of these biases on teleconnections should be improved. Increased vertical resolution appears to be necessary.

- QBO teleconnections show promise but are too weak; a more realistic QBO vertical structure is desirable, and may enhance predictability.

4.1.2 Strengths to be continued and developed

- ECMWF has a strong tradition of high quality multi-decadal ocean re-analyses and consistent real-time analyses. This needs to be maintained, with continued attention to ENSO skill, other tropical oceans and the global ocean state. Consistent real-time analyses need to be provided, and appropriate solutions need to be found to support both evolving medium-range needs and the stability needed by long-range forecast systems.
- Land surface initial conditions have received much attention in both S4 and SEAS5, but further enhancements are still possible in terms of offline surface re-analysis, and consistency between re-analyses and real-time analyses should continue to be pursued and enhanced.
- Evaluation of new model cycles and feedback to the model development teams have long been important at ECMWF, and become even more so with a seamless approach to the forecast model. Enhancing our diagnostic capabilities and improving our testing and feedback protocols is important. There is scope to benefit from collaborations with C3S and externally, both for forecast metrics and climate diagnostics.

4.1.3 Advancing the state of the art

- Radiatively interactive prognostic ozone is important for modelling the lower stratosphere, and is important for the QBO. This will be pursued as a pan-ECMWF effort, as discussed in the accompanying paper on atmospheric composition.
- Long-term changes in tropospheric aerosol need to be treated appropriately and consistently across ECMWF forecast systems – simple solutions should be effective initially, although for some species modelling may also be valuable. Volcanic aerosol remains a challenge that could become critically important at any time, and there are ideas on how the current approach could be improved.
- The increasing realism of our forecast systems and the apparent predictability “gap” in seasonal forecast systems for the NH are providing increasing incentives for general predictability research, to improve our understanding of processes, errors and priorities for model development. Research is also needed to address the probabilistic prediction of weather statistics and weather events – regimes, blocking episodes, heat waves. This problem is best suited for collaborative research, and currently we are actively seeking engagement with different institutions.

4.2 Seamless strategy

A “seamless” approach to modelling has several key benefits: it focusses our development and testing resources on a single model version; it helps guarantee that the long-range forecast model has the best possible representation of fast physics processes; and it ensures that benefits of improved slower processes are shared between long-range and extended-range forecasts and other IFS configurations. In particular, it is our goal that the IFS configuration will be identical for all time ranges from medium-range to seasonal, apart from horizontal resolution. SEAS5 has already taken major strides in being near-identical to the extended range system, which is in turn linked to the medium-range ensemble.

The seamless strategy means that all model changes motivated by the medium-range must be carefully assessed for their impact on the seasonal and extended ranges. Equally, changes important for longer timescales need assessment in the medium-range. Enhanced evaluation of model cycles at seasonal time scales becomes a key building block for the seamless strategy, with consideration of performance on seasonal timescales becoming part of the overall judgement on the merits of a cycle (independent of whether it is or is not planned to upgrade the operational seasonal system). Plans for this are discussed in the companion SAC paper on “Evaluating model cycles”. It should be noted that running the necessary forecast experiments is straightforward, but that evaluation is not. Significant technical development and substantial scientific work is needed to create efficient evaluation software, metrics and scorecards, necessary to allow timely feedback on model developments. This is an area of development where shared interests with the C3S seasonal activities can greatly help.

A further step, would be to move to an approach which updates the seasonal system with every operational implementation (as already done for the extended-range). The benefits would be to take advantage more quickly of scientific developments that could increase the skill of operational seasonal predictions. However, this would incur increased computational costs associated with the need to rerun the full set of reforecasts with each cycle (e.g. SEAS5 reforecasts estimated to have used 3% of the total HPC resource in 2017). User needs are also important, and users may be split on the benefits of frequent upgrades versus the benefits of stable systems. Whether to move as far as upgrading every cycle therefore remains an open question but, on balance, our judgement is that we should aspire, as a minimum, to more frequent upgrades of the operational seasonal system.

Within the context of a possible multi-resolution medium and extended range ensemble, it is possible to consider part of the ensemble covering seasonal ranges, providing the basis for a unified set of ensemble products covering all timescales. Such an approach could be configured in a variety of ways. One important variation is a lag-average approach to seasonal forecasting, which we plan to explore both with and without a multi-ensemble framework.

Finally, the details of the implementation of the seamless approach will always be pragmatic rather than dogmatic. We choose to follow a seamless strategy because of the overall benefits (for performance and in making most efficient uses of resources) that we believe come (for all timescales) through developing a system suitable for use from medium-range to seasonal. Too much divergence from the single, unified system approach can quickly lose some of the advantages, but that need not mean that minor (and hopefully) temporary divergences cannot be considered.

4.3 Roadmap towards SEAS6

There are several steps needed to obtain a seamless modelling system. The first is to unify vertical resolution. This has important benefits for model tuning and assessment, and moving the ensemble configurations from L91 to L137 is now known to give significant benefits in terms of stratospheric biases. Indeed, it may be that vertical resolution is increased slightly beyond L137. Since the seasonal forecast configuration is only a small fraction of the cost of the medium-range systems, the decision making will be largely driven by cost-benefit considerations in the medium-range; nonetheless, seasonal-length integrations will help inform the debate.

There are some minor non-seamless details to sort out. We want to unify the treatment of sulphate aerosol and other anthropogenic aerosols, to allow decadal variability in the re-forecasts while maintaining compatibility with the real-time forecast system using the latest CAMS-based climatology. Such a unification will in principle benefit the ENS re-forecasts: the impact is expected to be negligible in the medium-range, but might have a marginal impact at the extended range and for the EFI. It is also hoped that we can move to a common linearized ozone scheme and radiatively interactive ozone.

A major ingredient of SEAS6 is expected to be a new ocean re-analysis, driven by ERA5 and without the North Atlantic problem. We also seek to reduce the sea-ice biases, especially those associated with the fast processes -the model is sluggish creating or meting sea-ice during the transition seasons. A switch to a new sea-ice model, or improved IFS behaviour over sea-ice, might allow a reduction of these sea-ice biases. Anticipated stratospheric improvements will feed into the implementation in SEAS6. The longer time-period covered by the extended ERA5 should allow a longer-term ocean re-analysis and potentially a longer re-forecast period; the costs and benefits of the latter will need to be assessed. We also expect that the re-forecast land initial conditions will be produced with a land surface assimilation system.

The likely time-frame for an operational release of SEAS6 is 2021/22. The key limiting factors in this are the time taken to develop and then produce a new ocean re-analysis, and the need to work around the move of HPC facility to Bologna. Strategically, our preference is for more frequent upgrades of the seasonal configuration, to keep it always close to the extended-range configuration, but it is hard to envisage an earlier date for SEAS6 given the need for a new ocean re-analysis and HPC constraints.

Acknowledgements

Nils Wedi, Irina Sandu, Richard Forbes, Anca Brookshaw, Erik Andersson, Andy Brown and Florence Rabier provided helpful comments on earlier drafts of this paper. We also acknowledge the work of Simona Briceag and Axel Bonet in setting up the operational suites for SEAS5, and all other ECMWF staff who have actively contributed to the development of the IFS model and the successful implementation of SEAS5.

This study has been partially funded by the Copernicus Climate Change Service. ECMWF implements this Service and the Copernicus Atmosphere Monitoring Service on behalf of the European Commission. Some of the diagnostics presented were developed under the auspices of the H2020 project AtlantOS.

References

Anderson, D., T. Stockdale, M. Balmaseda, L. Ferranti, F. Vitart, P. Doblas-Reyes, R. Hagedorn, T. Jung, A. Vidard, A. Troccoli and T. Palmer, 2003: Comparison of the ECMWF seasonal forecast Systems 1 and 2, including the relative performance for the 1997/8 El Nino. ECMWF Tech Memo 404.

Anderson, D., T. Stockdale, M. Balmaseda, L. Ferranti, F. Vitart, F. Molteni, F. Doblas-Reyes, K. Mogenson and A. Vidard, 2007: Development of the ECMWF seasonal forecast System 3. ECMWF Tech Memo 503.

Barnston, A.G., M.K. Tippett, M.L. L'Heureux, S. Li and D.G. DeWitt, 2012: Skill of Real-Time Seasonal ENSO Model Predictions during 2002–11: Is Our Capability Increasing? Bull. Amer. Meteor. Soc., 93, 631–651, https://doi.org/10.1175/BAMS-D-11-00111.1

Balmaseda, M.A., K. Mogensen and A.T. Weaver, 2013: Evaluation of the ECMWF ocean reanalysis system ORAS4, Q.J.R. Meteorol. Soc, 139, 1132–1161.

Booth, J.F., L. Thompson, J. Patoux and K.A. Kelly, 2012: Sensitivity of midlatitude storm intensification to perturbations in the sea surface temperature near the Gulf Stream. Mon. Weather Rev., 140, 1241–1256.

Breivik, Ø., K. Mogensen, J.-R. Bidlot, A. Balmaseda and P.A.E.M. Janssen, 2015: Surface wave effects in the NEMO ocean model: Forced and coupled experiments. J. Geophys. Res. (Oceans), 120, 2973–2992.

Bryan, Frank O. et al., 2010: Frontal scale air–sea interaction in high-resolution coupled climate models. J. Clim. 23, 6277-6291.

Buckley, M.W. and J. Marshall, 2016: Observations, inferences, and mechanisms of the Atlantic Meridional Overturning Circulation: A review. Reviews of Geophysics, 54(1), 5–63.

Cassou, C., 2008: Intraseasonal interaction between the Madden–Julian Oscillation and the North Atlantic Oscillation. Nature 455: 523–527.

Chassignet, E.P. and D.P. Marshall, 2008: Gulf Stream separation in numerical ocean models. In M. W. Hecht & H. Hasumi (Eds.), Ocean Modeling in an Eddying Regime (pp. 39–61). American Geophysical Union (AGU).

Czaja, A. and C. Frankignoul, 2002: Observed impact of Atlantic SST anomalies on the North Atlantic Oscillation. J. Clim., 15, 606–623.

Davini, P., J. von Hardenberg, S. Corti, A. Subramanian, H. Christensen, S. Juricke, P.A.G Watson, A. Weisheimer and T.N. Palmer, 2017: Climate SPHINX: evaluating the impact of resolution and stochastic physics parametrizations in climate simulations. Geosci. Model Dev., doi:10.5194/gmd-10-1383-2017.

Dee, D. P., S.M. Uppala, A.J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M.A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A.C.M. Beljaars, L.V.D. Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A.J. Geer, D.P. Dee, L. van de Berg, L. Haimberger, S.B. Healy, H. Hersbach, E.V. Hólm, I. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A.P. Mcnally, B.M. Monge-Sanz, J.-J. Morcrette, B.K. Park, C. Peubey, P. de Rosnay, C. Tavolato, C., J.-N. Thépaut and F. Vitart, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q.J.R. Meteorol. Soc., 137, 553–597.

Duan, W. and J. Hu, 2016: The initial errors that induce a significant "spring predictability barrier" for El Niño events and their implications for target observation: Results from an earth system model. Climate Dynamics, 46, 3599–3615.

Good, S.A., M.J. Martin and N.A. Rayner, 2013: EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates*,* J. Geophys.l Res. (Oceans*)*, 118, 6704–6716.

Graham, Tim, 2014: The importance of eddy permitting model resolution for simulation of the heat budget of tropical instability waves. Ocean Modelling**,** **79**, 21-32.

Hewitt, Helene T. et al, 2016: The impact of resolving the Rossby radius at mid-latitudes in the ocean: Results from a high-resolution version of the Met Office GC2 coupled model. Geosci. Model Dev. 9, 3655.

Hogan, R., M. Ahlgrimm, G. Balsamo, A. Beljaars, P. Berrisford, A. Bozzo, F. Di Giuseppe, R.M. Forbes, T. Haiden, S. Lang, M. Mayer, I. Polichtchouk, I. Sandu, F. Vitart and N. Wedi, 2017: Radiation in numerical weather prediction. ECMWF Technical Memorandum 816.

IFS Documentation 43R1, ECMWF, 2016.

Johns, William E. et al, 2011: Continuous, array-based estimates of Atlantic Ocean heat transport at 26.5 N. J. Clim., 24, 2429-2449.

Johnson, S. J., Tim Stockdale, Laura Ferranti, Magdalena Balmaseda, Franco Molteni, Linus Magnusson, Steffen Tietsche, Damien Decremer, Antje Weisheimer, Gianpaolo Balsamo, Sarah Keeley, Kristian Mogensen, Hao Zuo and Beatriz Monge-Sanz, 2018: SEAS5: The new ECMWF seasonal forecast system. Geosci. Model Dev.*,* submitted.

Leutbecher, M., S.-J. Lock, P. Ollinaho, S. Lang, G. Balsamo, P. bechtold, M. Bonavita, H. Christensen, M. Diamantakis, E. Dutra, S. English, M. Fisher, R. Forbes, J. Goddard, T. Haiden, R. Hogan, S. Juricke, H. Lawrence, D. MacLeod, L. Magnusson, S. Malardel, S. Massart, I. Sandu, P. Smolarkiewicz, A. Subramanian, F. Vitart, N. Wedi and A. Weisheimer, 2017: Stochastic representations of model uncertainties at ECMWF: State of the art and future vision*.* Q.J.R. Meteorol. Soc.*,* doi:10.1002/qj.3094.

Lin, H., G. Brunet and J. Derome, 2009: An observed connection between the North Atlantic Oscillation and the Madden–Julian Oscillation. J. Climate, 22, 364–380.

Molteni, F., T. Stockdale, M. Balmaseda, G. Balsamo, R. Buizza, L. Ferranti, L. Magnusson, K. Mogensen, T. Palmer and F. Vitart, 2011: The new ECMWF seasonal forecast system (System 4), ECMWF Tech Memo 656.

Molteni F., T. Stockdale and F. Vitart, 2015: Understanding and modelling extra-tropical teleconnections with the Indo-Pacific region during the northern winter. Climate Dyn*.,* 45, 3119-3140.

Palmer, T.N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G.J. Shutts, M. Steinheimer and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. ECMWF Tech. Memo. 598.

Palmer T.N., 2012: Towards the probabilistic Earth-system simulator: a vision for the future of climate and weather prediction. Q. J. R. Meteorol. Soc. 138, 841–861.

Parfitt, R., A. Czaja, S. Minobe and A. Kuwano-Yoshida, 2016: The atmospheric frontal response to SST perturbations in the Gulf Stream region. Geophys. Res. Lett., 43, 2299–2306.

Polichtchouk, I., R.J. Hogan, T.G. Shepherd, P. Bechtold, T. Stockdale, S. Malardel, S.-J. Lock, S.-J. and Magnusson, L., 2017: What influences the middle atmosphere circulation in the IFS? ECMWF Technical Memorandum 809.

Roberts, C.D, R. Senan, F. Molteni, S. Boussetta, M. Mayer and S. Keeley, 2018: Climate model configurations of the ECMWF Integrated Forecast System (ECMWF-IFS cycle 43r1) for HighResMIP. Geosci. Model Dev., submitted.

Roberts, Malcolm J. et al, 2016: Impact of ocean resolution on coupled air-sea fluxes and large-scale climate. Geophy.s Res. Lett. 43**,** 10,430-10,438.

Robson, J., R. Sutton and D. Smith, 2013: Predictable Climate Impacts of the Decadal Changes in the Oceans in the 1990s. J. Climate*,* 26, 6329–6339.

Sardeshmukh, P.D. and B.J. Hoskins, 1988: The generation of global rotational flow by steady idealized tropical divergence. J. Atmos. Sci., 45, 1228-1251.

Scaife, Adam A. et al, 2011: Improved Atlantic winter blocking in a climate model. Geophys. Res. Lett. 38**.**

Scaife, Adam A. et al 2018: Tropical Rainfall Predictions from Multiple Seasonal Forecast Systems. International Journal of Climatology. Accepted.

Shepherd, T.G., I. Polichtchouk, R.J. Hogan and A.J. Simmons, 2018: Report on Stratosphere Task Force. ECMWF Technical Memorandum 824.

Simpson, I.R., J.T. Bacmeister, I. Sandu, and M.J. Rodwell, 2018: Why Do Modeled and Observed Surface Wind Stress Climatologies Differ in the Trade Wind Regions? J. Climate, 31, 491–513, https://doi.org/10.1175/JCLI-D-17-0255.1

Shonk, J.K.P., E. Guilyardi, T. Toniazzo, S.J. Woolnough and T. Stockdale, 2018: Identifying causes of Western Pacific ITCZ drift in ECMWF System 4 hindcasts. Clim Dyn, 50, 939-954.

Shutts, G.J., 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. Q.J.R. Meteorol. Soc. 131, 3079–3102.

Stockdale, T.N., D.L.T. Anderson, M.A. Balmaseda et al, 2011: ECMWF seasonal forecast system 3 and its prediction of sea surface temperature. Clim Dyn 37, 455.

Tibaldi, Stefano and Franco Molteni, 1990: On the operational predictability of blocking. Tellus A: Dynamic Meteorology and Oceanography 42.3**,** 343-365.

Tietsche, S., M.A. Balmaseda, H. Zuo and K. Mogensen, K., 2017: Arctic sea ice in the global eddy-permitting ocean reanalysis ORAP5*,* Climate Dynamics, 49**,** 775–789.

Vitart, F., J.L. Anderson and W.F. Stern, 1997: Simulation of interannual variability of tropical storm frequency in an ensemble of GCM integrations. J. Climate, 10, 745-760.

Vitart, F. and T.N. Stockdale, 2001: Seasonal forecasting of tropical storms using coupled GCM integrations*.* Mon. Wea. Rev, 129(10), 2521-2527.

Vitart, F., M.A. Balmaseda, L. Ferranti and D. Anderson, 2003: Westerly wind events and the 1997/98 El-Niño event in the ECMWF seasonal forecasting system, J. Clim.,16, 3153–3170.

Weisheimer, A., S. Corti, T.N. Palmer and F. Vitart, 2014: Addressing model error through atmospheric stochastic physical parametrizations: Impact on the coupled ECMWF seasonal forecasting system. Phil. Trans. R. Soc. *A*, 372,201820130290, doi: 10.1098/rsta.2013.0290.

Weisheimer, A., N. Schaller, C. O'Reilly, D. MacLeod and T.N. Palmer, 2017: Atmospheric seasonal forecasts of the 20th Century: multi-decadal variability in predictive skill of the winter North Atlantic Oscillation and their potential value for extreme event attribution. Q.J.R. Meteorol. Soc., 143, 917-926.

Wheeler, M.C. and H.H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. Mon. Wea. Rev. 132(8), 1917–1932.

Wilks, D.S., 2011: *Statistical Methods in the Atmospheric Sciences*. Academic Press.

Zuo, H., M.A. Balmaseda, E.D. Boisseson, S. Hirahara, M. Chrust and P. de Rosnay, 2017: A generic ensemble generation scheme for data assimilation and ocean analysis, ECMWF technical memorandum 795.

Zuo, H., M.A. Balmaseda and K. Mogensen, 2017: The new eddy-permitting ORAP5 ocean reanalysis: description, evaluation and uncertainties in climate signals, Climate Dynamics, 49, 791–811.

Zuo, H., M.A. Balmaseda, K. Mogensen and S. Tietsche, 2018: OCEAN5: the ECMWF Ocean Reanalysis System ORAS5 and its Real-Time analysis component, ECMWF Technical Memorandum 823.

Appendix

A.1. A brief history of seasonal forecasting systems at ECMWF

System 1 (1997-2002) The first seasonal forecasting system (Stockdale et al, 1998) was a pilot, with many ad hoc technical features, such as the re-forecasts being based on “burst mode” ensembles and the real-time forecasts being based on lag-ensemble (a single-member forecasts run each day. The IFS cycle used was 15r8, resolution was T63 with 31 levels up to 10 hPa, and the ocean was a 2°x2° version of the HOPE ocean model with equatorial refinement. Ocean initial conditions were based on an OI assimilation system, taking advantage of then recent developments on atmospheric reanalyses (ERA-15) and ocean observing system (TAO array). The original set of re-forecasts covered only 1991-1996, although this was later extended, and all re-forecasts were eventually re-archived in MARS as operational data. Although this first system had a substantial cold bias in surface temperatures, in many ways it performed very well, capturing many of the anomalies across the world associated with the record-breaking El Niño of 1997/98 and then the strong La Niña conditions that followed in subsequent years.

System 2 (S2: 2002-2007): It was with System 2 that the seasonal forecasting become a proper end-to-end forecasting system, with specific design of the forecast suite, ensemble generation, archiving and products. S2 consisted of Cy23r4 of the IFS at TL95 resolution with 40 vertical levels. The HOPE ocean model now had a base resolution of 1°x1°, with equatorial refinement. Ocean initial conditions came from an assimilation system based on an OI analysis as before, but now extended to a 5-member ensemble analysis, using wind perturbations to help increase the spread and partially represent the uncertainty in the ocean sub-surface. A set of SST perturbations was used to represent uncertainty in ocean surface initial conditions. Re-forecasts and real-time forecasts were run with a consistent ensemble structure, with "burst" ensembles from the 1st of the month in all cases. Stochastic physics was used to assure a reasonable initial growth of spread in atmospheric fields. Atmosphere and land surface initial conditions came from a mixture of ERA15 and ECMWF operations. Re-forecasts were 1987-2001 (15 years) with 5 members per month. Data were archived consistently in the mars SEAS stream, for both the ocean and atmosphere models. Scientifically, the performance of S2 was somewhat disappointing, in that overall it did not lead to clearly better forecasts than its predecessor. This was discussed in a report to the SAC at the time (Anderson et al, 2003), the main conclusion of which was that while improvements in the ocean model and analyses had benefitted the system, the changes in the IFS between Cy15r8 and Cy23r4 had caused significant damage to the ENSO forecasts, due to an increase in easterly wind bias and more particularly to a sharp reduction in wind variability, resulting in substantially damped ENSO forecasts. The performance of System 2 is also documented in the peer-reviewed literature (Oldenburgh et al, 2004 a,b).

System 3 *(S3: 2007-2011)* was a major advance on the state of the art, and a step forward regarding an Earth system approach: time-varying CO_2 levels were included, necessary to properly account for anthropogenic warming. S3 employed Cy31r1 of the IFS at TL159 resolution with 62 levels, up to approximately 5 hPa. Major improvements were made to the ocean reanalyses, which now took advantage of ERA40 long records, including assimilation of salinity and altimetry data and an adaptive multivariate bias adjustment. The ocean resolution was unchanged from System 2, but modifications were made to the free surface solver to enable assimilation of altimeter sea-level. Wind and SST perturbations used to generate the ensemble of ocean initial conditions were revised. Atmosphere and land surface initial conditions came from a mixture of ERA40 and ECMWF operations. Singular vectors were used to perturb the atmosphere initial conditions, and stochastic physics continued to be used as in the medium-range ensemble. The forecast length was increased from 6 to 7 months, and quarterly forecasts out to 13 months were introduced, to give an ENSO outlook. The re-forecast set was increased to 25 years (1981-2005) and 11 members (41 members for Feb, May, Aug and Nov starts). Ocean and atmosphere data were both archived in the new MMSF streams in MARS, consistent with the development of multi-model seasonal forecasting at ECMWF. Full details are given in the published version of an SAC report (Anderson et al, 2007). S3 performed very well in terms of SST prediction and demonstrated improvements in model climate and the amplitude of ENSO variability, although it was still underactive.

System 4 (S4: 2011-2017) consisted of Cy36r4 of the IFS at TL255 resolution (80 km grid point resolution) with 91 levels, covering the whole of the stratosphere for the first time. The non-orographic gravity wave scheme in the IFS was re-tuned to produce good forecasts of the QBO. A treatment of time-varying volcanic aerosol was included via a simple namelist specification, designed to allow real-time use should the need arise. Radiatively interactive ozone was also activated, making use of the Cariolle scheme which has long been used in the IFS. Sea-ice was prescribed using a sampling of the 5 preceding years, allowing both trends and uncertainty in sea-ice to be represented. A major change was the ocean: the HOPE ocean model, no longer supported by an external community and lacking MPI parallelisation, was replaced by NEMO. The ORCA1 (1°x1°) configuration was used was comparable to the HOPE model it replaced, but the tripolar grid and netCDF output meant that ocean data could no longer be archived in MARS. The new ocean model was introduced with a new variational ocean assimilation system, which however included many of the previous features (a 5-member ensemble driven with wind perturbations, an adaptive multivariate bias adjustment). Atmosphere initial conditions came from a mixture of ERA Interim and ECMWF operations, while land surface conditions for the re-forecasts came from a specially produced offline run of the HTESSEL surface model, which was later released for general usage as the dataset "ERAI land". The re-forecast set increased to 30 years (1981-2010) and 15 members (51 members for Feb, May, Aug and Nov starts), and the real-time ensemble size increased to 51. S4 was described in a report to the SAC in 2011 (Molteni et al*, 2011*). The biggest challenge in implementing S4 was that developments in the IFS in the preceding years had led to a substantial increase in equatorial wind stress bias which drove a strong cooling of the equatorial Pacific. The increased wind bias was the result of changes which had greatly improved the model MJO, and the difficulty was reducing the bias without undoing the improved intra-seasonal variability. The size of the bias in cycles just before 36r4 was so large as to inflict serious damage on the ENSO forecasts, and consideration was given as to whether flux correction might be needed. However, Cy36r4 resulted in a slight decrease in the bias, which was enough to tip the coupled model back into a regime where ENSO forecasting worked well. The remaining biases were still strong, and resulted in an amplitude of ENSO variability that was substantially too large, but this latter was corrected in the "Nino plume" forecast products by a variance scaling. Despite these problems, S4 performed well in terms of a much-improved model climate globally, and improvements in a wide range of atmospheric forecast measures, both in terms of deterministic and probabilistic scores.

A.2. Treatment of volcanic aerosol in SEAS5

SEAS5 retains the S4 treatment of volcanic stratospheric sulphate aerosol, with damped persistence of an initially specified loading. The initial load for past dates is based on GISS data (2012 update[1] – a revised dataset compared to S4), with the horizontal distribution approximated by three numbers (northern hemisphere, tropical and southern hemisphere amounts) and the vertical distribution following a prescribed profile relative to the tropopause that is applied globally. The forecast is initialised using values from the month before the forecast starts via a namelist, and then evolved in time locally with damped persistence with an e-folding timescale of 400 days. SEAS5 cannot predict volcanic eruptions, but should a major eruption occur, manual estimates of the volcanic aerosol, based in part on Copernicus Atmosphere Monitoring Service (CAMS) SO2 analyses, would be included in the namelist to give a similar quality of representation as exists for major eruptions in the re-forecast period (El Chichon and Pinatubo). This treatment of volcanic aerosol is considered to allow a first order estimate of the scattering of visible light (which cools the surface, and depends only on the vertical integral of the aerosol), but is less satisfactory regarding the longwave-driven warming of the stratosphere, which

[1] https://data.giss.nasa.gov/modelforce/strataer/

depends on the details of both the horizontal and vertical distribution of aerosol and the time-evolution of droplet size, and drives changes in stratospheric winds which influence the NH winter circulation.

A.3. Niño plume variance re-scaling

The variance re-scaling that was necessary to give acceptable ENSO forecast skill in S4 has been deactivated for the Niño plume products in SEAS5. For much of the year the scaling would in any case make little difference, because the model variance already closely matches the observed variance. In forecasts verifying in March to May, however, particularly in the NINO3 region, variance scaling would make a substantial difference and give a substantial benefit to the r.m.s. error statistics and, more importantly, to the realism and accuracy of the real-time forecast products. *Figure A 1* shows this for forecasts from 1 December: the amplitude-corrected forecasts (blue) have substantially increased accuracy compared to the standard forecasts (red). The right hand of the figure shows the scaling applied, which amounts to a 40% reduction in anomaly amplitude for the fifth month (April). The amplitude ratio from the re-forecasts is made available to users of the web products as part of the verification information, and so users can in principle manually adjust the plotted Niño plumes to obtain more realistic forecast values when it is necessary to do so. However, this is inconvenient, and we suspect very few users will do this. On the other hand, the Niño plumes do represent what is happening in the model, even if they do not represent a credible forecast of the real world.

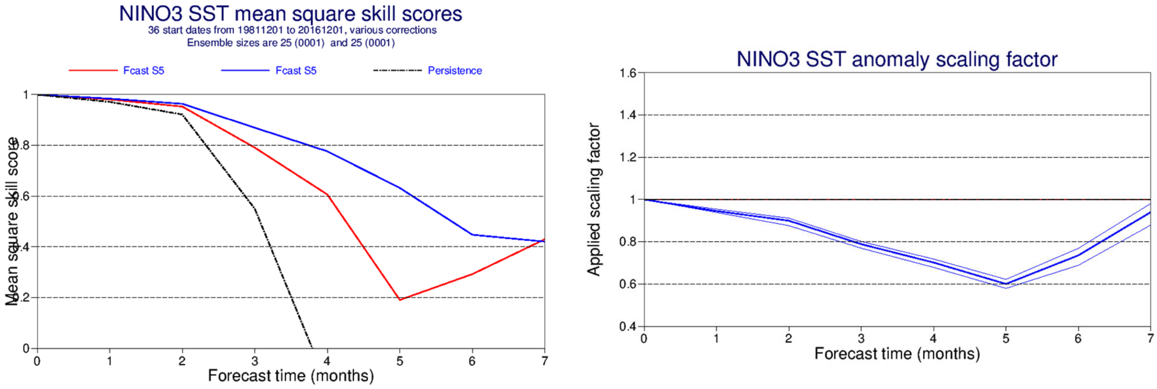


Figure A 1: SEAS5 NINO3 MSSS for forecasts starting on 1 December, for standard bias correction (red) and with additional variance scaling (blue). The scaling factor used is plotted as a function of lead time on the right, the thin lines showing the range of values resulting from the cross-validation procedure.

A.4. Evolution of ENSO Forecast skill: MAE metric

We compare operational forecasting systems for the available common period (1987-2002) and ensemble size (5 members), using a metric we have presented previously to the SAC (Anderson et al, 2007, page 24), namely the Mean Absolute Error (MAE) of forecasts for months 1-6 of three key ENSO SST anomaly indices (*Table A 1*). The Mean Absolute Error was chosen at the time to focus attention on "typical" forecast errors rather than being overly dominated by "busts" which were more frequent in the early systems. SEAS5 has the lowest error of all systems in all three NINO regions presented. S4 is an outlier due to it having a significant level of over-activity, and although S4 anomaly correlation was high, the excessive amplitude of anomalies hurt the MAE and RMSE scores. The excess anomaly amplitude in S4 was considered acceptable for two reasons: firstly, the ENSO plumes were plotted with the model variance re-scaled to the observed variance, which effectively removed the excess amplitude from view; and secondly, in terms of impact on the atmosphere and ENSO teleconnections, the excess amplitude appeared to do no harm. The second set of figures for S4 in the table show the scores after variance correction, and represent the skill of the published S4 forecasts. SEAS5 is a substantial advance on the variance-corrected S4 as well as the standard S4 score.

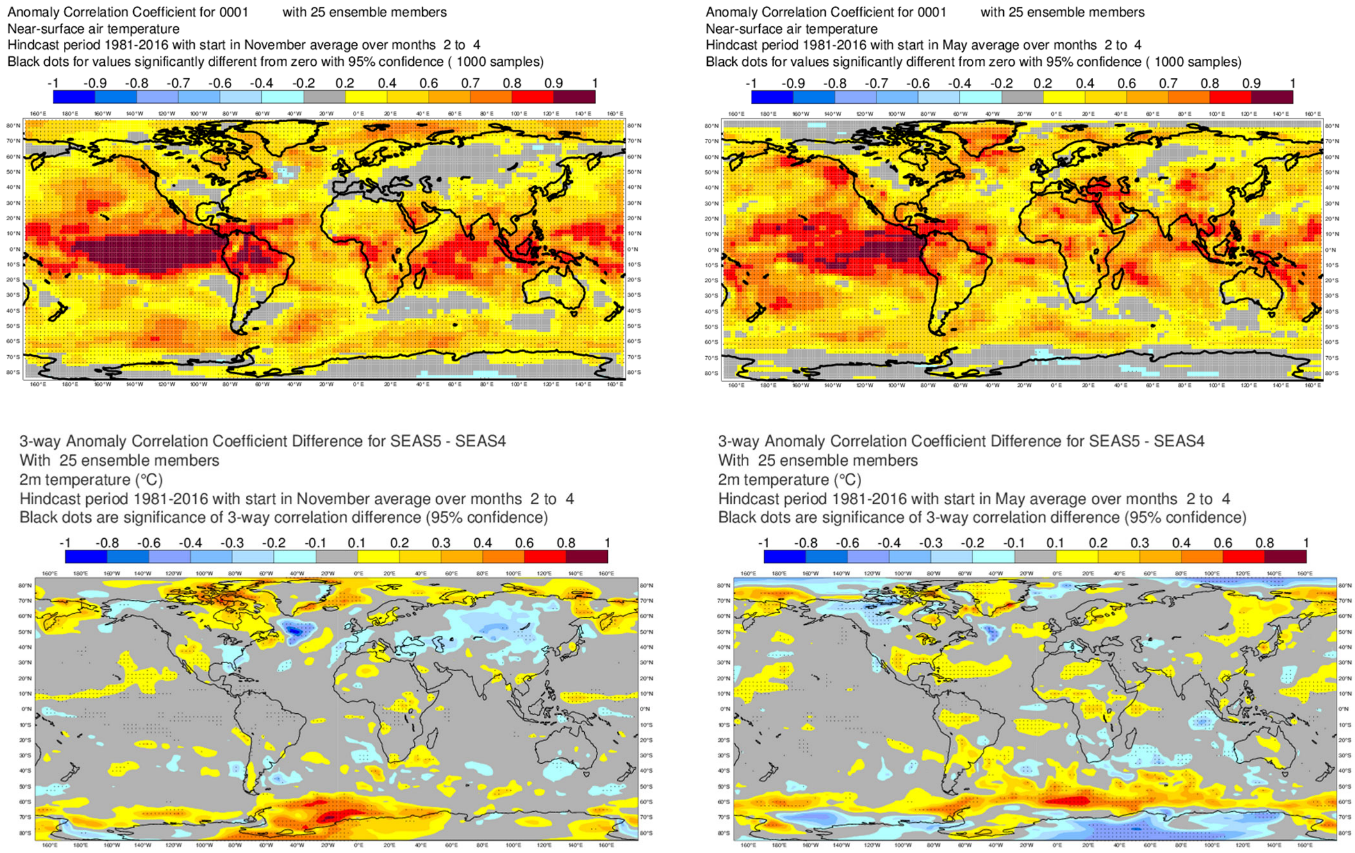
Table A 1: The Mean Absolute Error in SST, for 5-member ensemble mean forecasts from 192 start dates from the common period 1987 to 2002, for the five operational systems at ECMWF. Values are shown for three NINO regions, and a composite metric being the sum of the three regions. For S4, values for variance-scaled forecasts are also given.

Version	NINO3	NINO3.4	NINO4	SUM(N3+3.4+4)
SEAS5	0.340	0.304	0.248	0.892
S4	0.424 / 0.360	0.384 / 0.349	0.295 / 0.289	1.119 / 0.998
S3	0.374	0.332	0.262	0.968
S2	0.403	0.388	0.319	1.110
S1	0.454	0.428	0.279	1.161

A.5. Maps of anomaly correlation: T2m and precipitation

The local correlation between SEAS5 ensemble-mean and ERA-interim T2m for lead time 2-4 months is shown in the top panels of A2, for forecasts initialized in November (left) and May (right). An overall high level of skill for near-surface temperature is evident over the tropics and particularly over the oceans. Some extra-tropical land regions, depending on the season, also show a useful level of skill. There is skill across northern and central Europe for winter temperature. For summer temperature, we see significant skill over South-Eastern Europe and the Mediterranean. Since these regions have been subject to a long-term warming trend, it is likely that a substantial part of the skill is associated with the model ability to represent the warming trend rather than the year-to-year variability.

The SEAS5 correlation differences with S4 (bottom panels of Figure A2) are calculated with a 3-way significance test applied. They show an overall improvement in the DJF surface temperature predictions north of 60N and south of 60S. This improvement appears to be associated with the introduction of a dynamical sea-ice model. Some positive improvement is found in the tropical and sub-tropical Eastern Pacific reaching the West Coast of America. There is also a drop of skill over the north-west Atlantic, discussed in Section 3.2.7, and large degradation over Central Asia and Siberia. Some improvement for JJA predictions is found over Greenland and Eastern Siberia. A substantial enhancement of skill is found in the S Hemisphere at about 60S, again associated with the ice edge. However, the skill is reduced in the Arctic Ocean, which may be related to a bias in the melt time of Arctic ice (section 3.1.3). In neither DJF or JJA are there evidences of significant skill improvement over Europe. At longer time ranges (month 5-7) SEAS5 exhibits clearly enhanced skill over the tropical oceans in both MAM and SON, and over the Arctic in autumn (not shown).



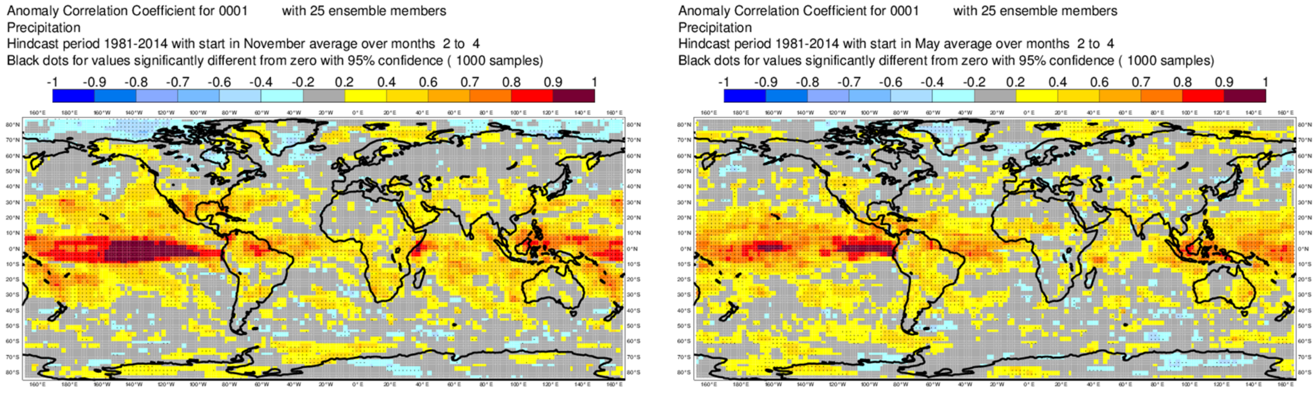
Anomaly Correlation Coefficient for 0001 with 25 ensemble members
Near-surface air temperature
Hindcast period 1981-2016 with start in November average over months 2 to 4
Black dots for values significantly different from zero with 95% confidence (1000 samples)

Anomaly Correlation Coefficient for 0001 with 25 ensemble members
Near-surface air temperature
Hindcast period 1981-2016 with start in May average over months 2 to 4
Black dots for values significantly different from zero with 95% confidence (1000 samples)

3-way Anomaly Correlation Coefficient Difference for SEAS5 - SEAS4
With 25 ensemble members
2m temperature (°C)
Hindcast period 1981-2016 with start in November average over months 2 to 4
Black dots are significance of 3-way correlation difference (95% confidence)

3-way Anomaly Correlation Coefficient Difference for SEAS5 - SEAS4
With 25 ensemble members
2m temperature (°C)
Hindcast period 1981-2016 with start in May average over months 2 to 4
Black dots are significance of 3-way correlation difference (95% confidence)

Figure A 2: Top) Anomaly correlation of the 1981-2016 SEAS5 T2m re-forecast verifying in DJF (left) and JJA (right), initialized in November and May respectively. Bottom) Difference between SEAS5 and S4 correlations Black dots indicate where correlation (or difference) differ from zero at 5% significance level.



Anomaly Correlation Coefficient for 0001 with 25 ensemble members
Precipitation
Hindcast period 1981-2014 with start in November average over months 2 to 4
Black dots for values significantly different from zero with 95% confidence (1000 samples)

Anomaly Correlation Coefficient for 0001 with 25 ensemble members
Precipitation
Hindcast period 1981-2014 with start in May average over months 2 to 4
Black dots for values significantly different from zero with 95% confidence (1000 samples)

Figure A 3: Anomaly correlation for 1918-2016 SEAS5 precipitation re-forecasts, verifying in DJF (left) and JJA (right) initialized in November and May respectively. Verified against GPCP

Skill for precipitation is generally lower and less spatially coherent than the skill for near-surface temperature (Figure A3). Precipitation is best predicted over parts of the tropical oceans, but seasonal prediction for rainfall over land is, with some exceptions, challenging. Although seasonal "local" (i.e. gridpoint) rainfall predictions are often not useful, average values on tropical regions have significant predictability and play a crucial role for extratropical predictability (Scaife et al 2018, Molteni et al. 2015). Differences in precipitation skill tend to be noisy and are not shown.

A.6. Aggregate anomaly correlation scores

One approach to the noisiness of skill difference maps is to define metrics and look at change in aggregate performance. Here we repeat and extend the analysis of anomaly correlation-based skill metrics that we presented in the SAC report on S4 (Molteni et al, 2011, Section 4.2).

For each calendar start date and lead time, we calculate the mean Fisher-z transformed temporal anomaly correlation for each point, based on a common period (1981-2005) and ensemble size (11 members). We take the area average over either the NHEX (poleward of 30N) or TR30 (extended tropics, 30N-30S) regions, inverse transforming back to anomaly correlation to obtain a single representative score for the region. Grid-point calculations are made after area-averaging fields to a common 2.5 degree grid. The 30° latitude line is chosen as the demarcation between the high predictability tropics/sub-tropics, and the lower predictability mid-latitudes.

For each NHEX aggregate score, we also calculate the sampling uncertainty due to the finite ensemble size. This is done with a basic bootstrap (not a percentile bootstrap) sampling over the ensemble members, and gives us a confidence interval for the *reproducibility* of the score, were we to repeat the experiment with different or larger ensembles. Because this uncertainty is related only to the ensemble sampling, it is independent between different experiments, so the uncertainty of differences between experiments can be estimated by standard methods of combining uncertainties. The uncertainties we find depend on the field, with precipitation having the lowest uncertainty, presumably due to it having the highest number of spatial degrees of freedom, and fields such as MSLP having higher uncertainty. We have not calculated uncertainty in TR30 scores in this way, but note that sampling errors are generally substantially smaller in the tropics due to the higher signal to noise ratio.

As well as plotting correlation differences as a function of start month (Section 3.1.2), we can also average the score differences across all lead times, giving the mean improvement in score for each variable and led time (Figure A4).

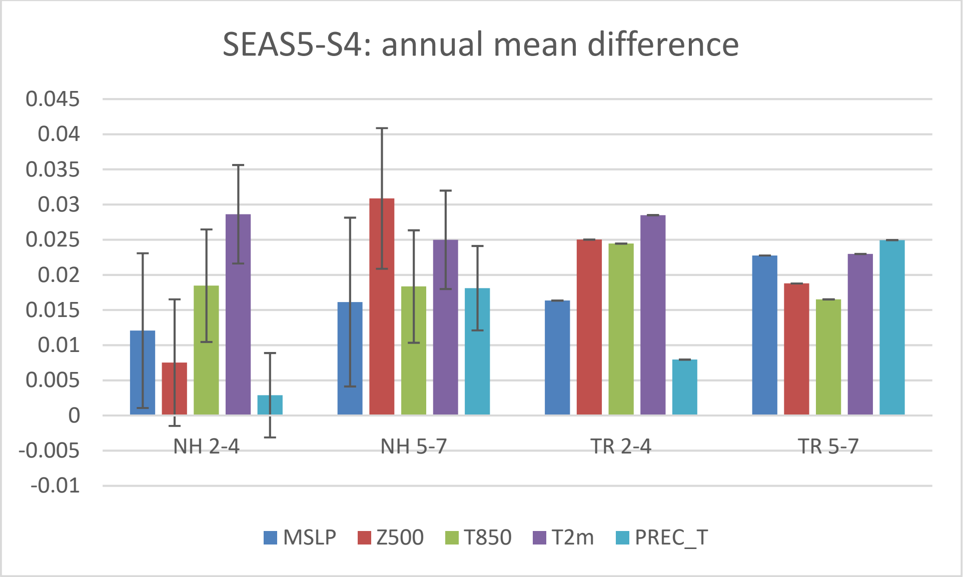


Figure A 4: Annual mean of SEAS5 – S4 differences in aggregated anomaly correlation over NHEX (NH) and TR30 (TR), for re-forecasts verifying at 2-4 month and 5-7 months, based on 1981-2010 15 member re-forecasts. Bars for NH indicate the 1 sigma sampling uncertainty in the correlation difference.

In all cases, the annual average gain in score is positive, although the difference is not always bigger than the sampling uncertainty, particularly for shorter lead times in the NHEX. At shorter leads, 2 metre temperature shows the strongest gains of all fields, and in NHEX it is only the temperature-related variables which show a statistically significant improvement. This is consistent with improved treatment of surface conditions (sea-ice, lakes, soil moisture) being an important contributor to the higher SEAS5 scores.

Finally, we note that all sampling uncertainties considered here relate to the uncertainty in ensemble mean due to limited ensemble size, and tell us the uncertainty in score differences *when assessing the 30 years being compared*. There is a further important uncertainty, namely the extent to which this period is representative of expected future behaviour. If we assume the climate system and forecast system are statistically stationary, and forecast errors from successive years are independent, we can calculate the sampling uncertainty of scores due to the finite length of our test period. This uncertainty has been calculated, and is typically several times larger than that from ensemble sampling. More relevant to comparing systems is the uncertainty in score differences, which will increase by much less than this, in the same way that 3-way tests are needed to assess correlation differences. Our software needs further development to enable uncertainty estimates of aggregate correlation differences to be calculated, so this uncertainty cannot be included in this analysis.

A.7. CRPSS and scorecards

Maps for CRPSS for SEAS5 and differences w.r.t. S4 are given in *Figure A 5*, for the 2-4 months forecast range, initialized in November (left) and May (right). Values above 20% are confined to the tropics. Precipitation forecasts are more skilful for DJF than for JJA, when the seasonal forecasts appear less skilful than climatology. There are significant differences between SEAS5 and S4, especially over the tropical Atlantic, where SEAS5 is better.

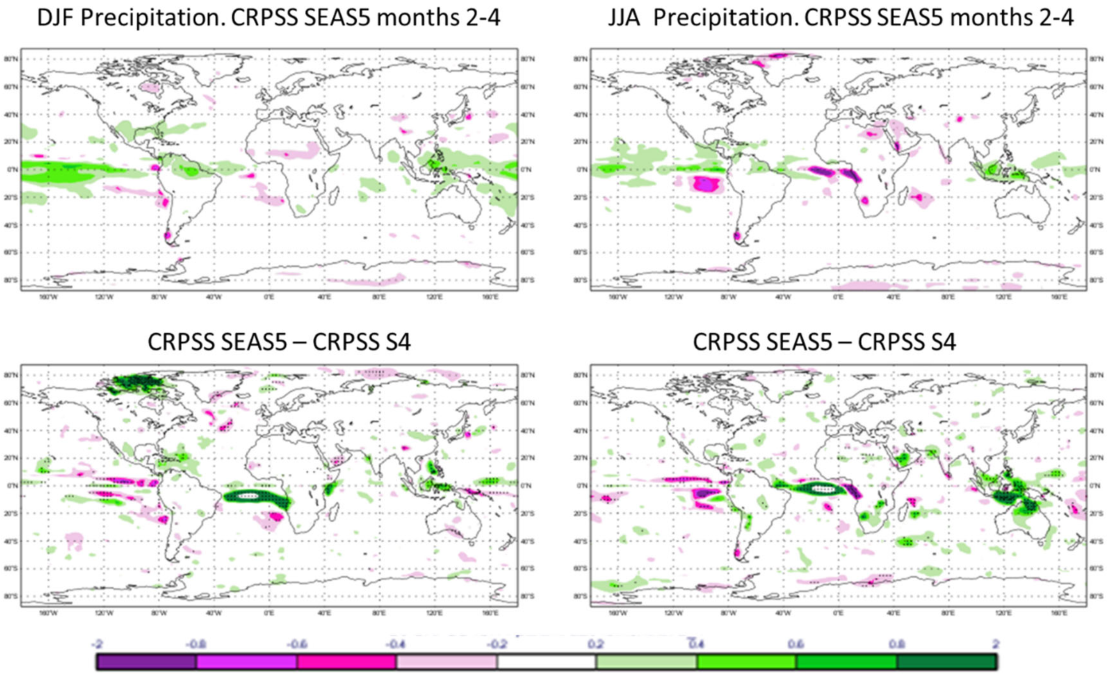

DJF Precipitation. CRPSS SEAS5 months 2-4
JJA Precipitation. CRPSS SEAS5 months 2-4
CRPSS SEAS5 – CRPSS S4
CRPSS SEAS5 – CRPSS S4
-2
-0.8
-0.6
-0.4
-0.2
0.2
0.4
0.6
0.8
2

Figure A 5: CRPSS of SEAS5 precipitation relative to climatology, for DJF (left) and JJA (right), with differences from S4 on the lower row. Scores are calculated from 25 member ensembles over the period 1981-2016, using months 2-4 of the May/November starts. Stippling in the lower plots represents differences significant at 5%, green is where SEAS5 is better than S4.

By aggregating CRPSS scores relative to climate over regions, and then differencing the scores of two systems, we can form scorecards, an example of which is shown in *Figure A 6*. The size of the score difference is indicated by the size of the circle, and significance at a 5% level is indicated with darker colours. Due to sampling limitations, statistical significance is hard to achieve.

This scorecard shows that SEAS5 exhibits significant improvement for precipitation for a number of tropical regions, notably Africa. Changes in T2m scores do not pass the significance test, but are consistently positive in South America, and show a sizeable apparent improvement over North America in DJF. On the other hand, there is an apparent deterioration in T2m score over Asia in winter.

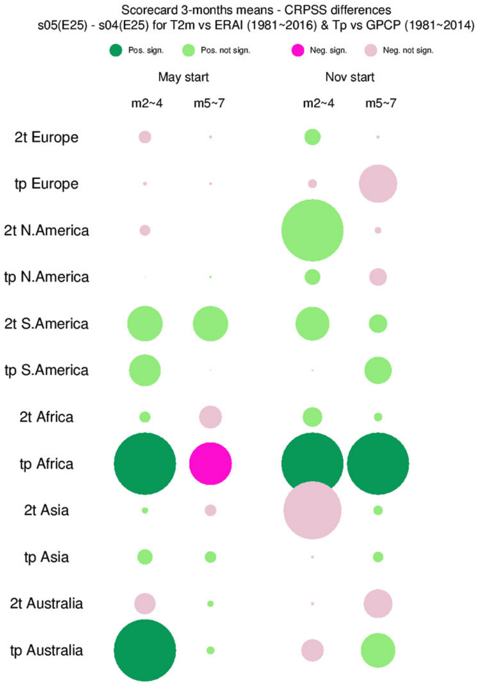


Figure A 6: Prototype scorecard for SEAS5 relative to S4, based on difference of CRPSS scores. The largest circles represent a skill difference of approximately 5%

A scorecard of this sort can summarize the information seen in the spatial maps, and makes it possible to show results from a larger number of fields and seasons on a single plot. However, it does not give the spatial information that can sometimes be visible in the maps. Particularly for T2m, changes in skill are often associated with changes in land surface or ocean processes which are highly localised, and global or regional metrics do not capture this. On the other hand, as shown with the anomaly correlation metrics discussed earlier, aggregating scores can greatly increase the effective sample size, and allow a more precise and statistically powerful test of whether a system has improved. In the end, a trade-off between spatial and physical detail and statistical robustness in constructing metrics and scores is inevitable. How to design the most appropriate, helpful and robust set of metrics for seasonal forecasting systems is still an active area of debate and research, and has been identified as a priority area for next year’s work plan.

A.8. Reliability scores

Reliability measures the ability of a forecast system to represent the observed frequency of events. If the forecast always provides the climatological probability, the system is in principle reliable, but has no skill. Nonetheless, reliability is an important requirement for issuing useful probabilistic predictions: a user should be able to trust a forecasting system with limited skill if the system is statistically reliable. Figure A 7 shows the reliability scores for T2m in the upper tercile over Europe (left) and the tropics (right). The reliability skill scores plotted here are aggregated over all grid points in the region over the whole reforecast period for S5 (1981-2016). Reliability skill scores are computed for warm events, three months average forecast anomalies in the upper third of the model climate distribution. The distribution of points in the figures clearly illustrates the difference in predictability of interannual variability over Europe and the tropics, with most forecasts in the European region being only moderately perturbed from climatological probabilities. Despite the limited signal over Europe, SEAS5 still shows a good level of reliability. In the tropics reliability is similarly good, but not perfect: the surface temperature forecasts are slightly over-confident with a systematic discrepancy between the forecast probabilities and observed frequencies. This is consistent with the model ENSO forecasts also being under-dispersive. In both the tropics and mid-latitudes, the level of reliability in SEAS5 is very similar to that seen in S4 (not shown).

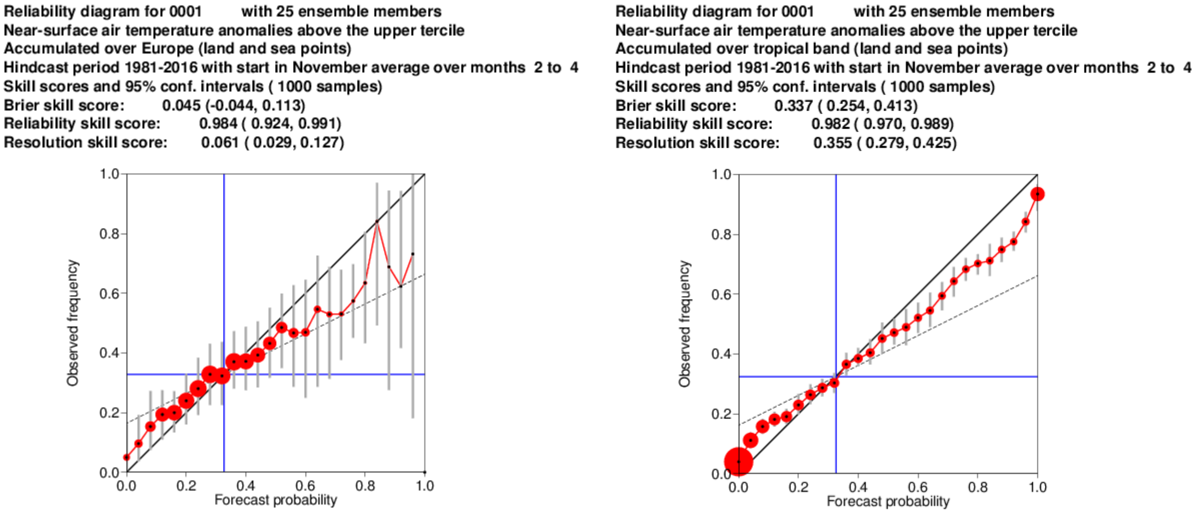


Figure A 7 SEAS5 reliability plots for T2m forecasts over Europe (left) and the tropics (right). The size of the dots reflects the number of forecast values in each probability bin.

A.9. Atmospheric circulation indices: NAO and PNA

Table A 2: Correlation for 1 November forecasts of DJF indices, from 25 member ensemble forecasts for 1981-2016, showing 95% confidence intervals based on re-sampling ensemble members. Also shown are results from a 51 member ensemble for S4 and SEAS5.

	NAO correlation	NAO (51 mem)	PNA correlation
S4	0.45 (0.35-0.66)	0.41(0.29-0.60)	0.72 (0.68-0.80)
SEAS5	0.32 (0.19-0.53)	0.43(0.33-0.59)	0.71 (0.67-0.77)
SEAS5 obs SST	0.41 (0.31-0.56)		0.71 (0.67-0.76)
SEAS5 L137	0.40 (0.29-0.57)		0.74 (0.71-0.79)

Table A 2 shows NAO and PNA scores for DJF forecasts from the 1st November, calculated using the standard ECMWF EOF-based definitions. Based on 25 member re-forecasts, SEAS5 scores lower than S4 for the NAO, but the difference is within sampling error. However, the extension of the SEAS5 re-forecasts is now complete for the November starts, so we can also compare scores based on the larger ensemble size. In this case, the best estimate we can give, SEAS5 is a whisker ahead with a correlation of 0.43. However, the error bars remain large, and all scores are indistinguishable from each other.

We also include results from SEAS5 run with observed SSTs, and for SEAS5-L137, for which a 25 member ensemble is available. Although in both cases the value of the mean correlation is higher than SEAS5, results are statistically indistinguishable from SEAS5 for both PNA and NAO. This suggests errors in SST (e.g. the North Atlantic problem discussed in section 3.2.7) are not having a dramatic negative impact on the scores of these indices, but the uncertainty is so large that nothing can be said about possible moderate impacts.

Sampling uncertainty has two sources. One is that the ensemble mean is uncertain due to a limited ensemble size for each forecast. For very noisy quantities such as the model NAO, this is a significant problem, and the confidence interval quoted in the table relates to this. It is relevant to the simple question of reproducibility, important when comparing experiments made for a fixed period. The second uncertainty is as to whether the 36-year period used is representative of the “longer term”. This uncertainty is even larger, and although we can estimate it using re-sampling methods, giving even wider error bars (e.g. Stockdale et al, 2015), a non-stationary climate makes it hard to be confident in what to expect for the future.

Previous experience has shown that NAO scores are sensitive to exactly how the NAO is defined. Operationally, ECMWF calculates NAO and PNA indices by projection of model Z500 fields onto EOFs derived from ERAI monthly mean variability in the winter period (DJFM), which is the method used in this analysis. As an example of an alternative definition, we follow Dunstone et al (2016) and calculate the NAO index as the MSLP difference between two small regions in the North Atlantic (Iceland: 25◦ to 20◦ W, 63◦ to 70◦ N and Azores: 28◦ to 20◦ W, 36◦ to 40◦ N). In this case we obtain an anomaly correlation of 0.30 in SEAS5 and 0.39 in S4 when using the 25-member ensemble, rather similar to the EOF method in this case.

The PNA scores have less sampling uncertainty (due to the larger signal/noise ratio in the PNA), and the model scores are very similar. In particular, there is very little difference between S4 and SEAS5, despite the large improvements in the MSLP bias in the north Pacific in SEAS5 that we show in A13.1.

A.10. Sea ice and Arctic 2m temperature forecasts

A substantial fraction of the Earth’s surface north of 70N is covered by sea ice, so it should be expected that near-surface air temperature co-varies with sea-ice extent. Figure A8 shows time series of the area mean of 2m temperature north of 70N for the same three-month ASO period from ERA-Interim, SEAS4 and SEAS5. Comparison of trend and interannual variability between the time series of *Figure 7*and A8 confirms that there is a substantial co-variability between sea-ice extent and Arctic near-surface air temperatures. This is especially striking in years with strong anomalies, such as 1996, 2007, and 2012. In ASO 1996, sea-ice extent jumped up by 1 Million km^2 w.r.t. the previous year, which was extremely well forecast by SEAS5, but not at all by S4. Likewise, 2m temperatures in ASO 1996 over the Arctic were 2K colder than in the previous year, and SEAS5 performs much better than S4 in forecasting this. In the years 2007 and 2012, sea-ice extent in the Arctic plummeted to unprecedented lows – these anomalies were again very well predicted by SEAS5, but not at all by S4. Corresponding warm anomalies in 2m temperature are forecast better in SEAS5 than in S4.

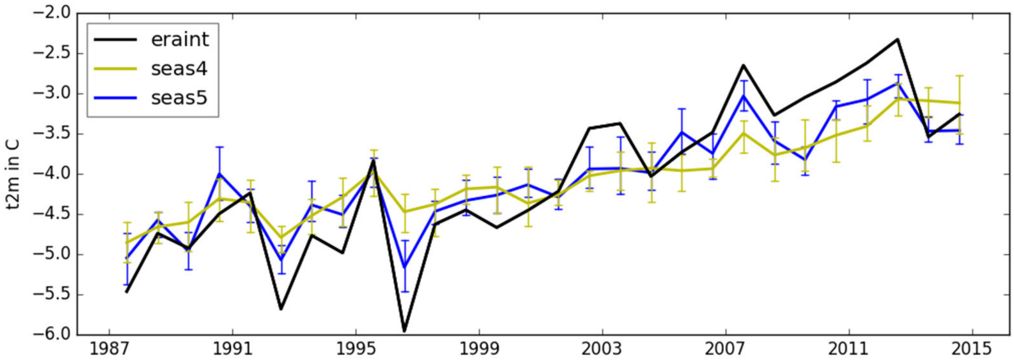


Figure A 8: Area mean of 2m temperature north of 70N during the sea-ice-minimum season ASO from ERA-Interim reanalysis (black), and in bias-corrected forecasts started from 1st July (yellow S4, blue SEAS5). The solid coloured lines connect the forecast ensemble means, and the error bars indicate the interquartile range of the ensemble.

A.11. IFS model upgrades from Cy36r4 to 43r1.

Table A 3 Changes in IFS between S4 (Cy36r4) and SEAS5 (43r1). (extracted from https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model).

Cycle	Date	Significant changes impacting model climate
37r2	18 May 2011	Improvements to cloud scheme formulation
37r3	15 Nov 2011	Modification of the entrainment/detrainment of convection Modification of the supersaturation and deposition rate for clouds Modification of the surface roughness
38r1	19 June 2012	Modified convective downdraught entrainment Changes to cloud ice fall speed, ice supersaturation, melting of ice to rain and freezing timescale of rain De-aliasing of the pressure gradient term, reducing numerical noise, allowing a reduction of the horizontal diffusion in the forecast Use of a new surface reanalysis to initialize the surface fields in the EPS re-forecasts
38r2	26 June 2013	Modified test parcel entrainment in boundary layer and shallow convection Adjustment of non-orographic gravity wave drag to be consistent with System-4 Oxygen absorption (radiative transfer)
40r1	19 Nov 2013	Convection - major change in diurnal cycle over land Vertical diffusion in stable conditions Orographic drag revisions Increase in the surface - atmosphere coupling for forest areas Upgrade to Nemo3.4 ocean model

41r1	12 May 2015	New surface climate fields (land-sea mask, sub-grid orography), also affecting number of land and sea points. New CO2/O3/CH4 climatologies from latest MACC-II reanalysis produced at ECMWF. Revised semi-Lagrangian extrapolation reducing stratospheric noise. Revised interpolation of moist variables in the upper-troposphere/lower stratosphere (UTLS). Cloud scheme change of rain evaporation, auto-conversion/accretion, riming, precipitation fraction. Improved representation of supercooled "freezing" rain. Modified convective detrainment. Activation of the lake model (FLAKE). Active use of wave modified stress in coupled mode. Revised sea-ice minimum threshold and sea-ice roughness length
41r2	8 Mar 2016	Improved representation of radiation-surface interactions with approximate updates every timestep on the full resolution grid leads to a reduction in 2m temperature errors near coastlines. Included surface-tiling for long-wave radiation interactions to reduce occasional too cold 2m temperature errors over snow. Improved freezing rain physics and an additional diagnostic for freezing rain accumulation during the forecast. Introduced resolution dependence in the parametrization of non-orographic gravity wave drag, reducing with resolution and improving upper stratospheric wind and temperature for HRES and ENS. Changed the parcel perturbation for deep convection to be proportional to the surface fluxes, reducing overdeepening in tropical cyclones. Increased cloud erosion rate when convection is active, to reduce cloud cover slightly and improve radiation, particularly over the ocean.
43r1	22 Nov 2016	A new CAMS ozone climatology is now used, consisting of monthly means of a re-analysis of atmospheric constituents (CAMSiRA) for the period 2003 to 2014. Changes to boundary layer cloud for marine stratocumulus and at high latitudes. Modifications to surface coupling for 2 metre temperature.

A.12. Summary of climate of IFS cycles

Here we report on some aspects of the evolution of IFS climate, as described section 3.2.1. Table A4 below lists the mean error (ME) and mean absolute errors (MAE) of the spatial bias patterns for SST, temperature at 500 hPa, TOA outgoing longwave radiation and TOA net incoming solar radiation. The SST is compared to the SST used in ERA-Interim, T500 to ERA-Interim and the TOA radiations against CERES satellite product available for the year 2000. The equivalent values for JJA are in Table A5. A graphical representation of the MAE values in the tables is given in *Figure A 9*.

The model troposphere has become significantly warmer due to the model changes since S4, which for DJF lead to a considerable decrease in the bias. For JJA, the tropospheric bias is also reduced, but recent model cycles have become too warm compared to ERA-Interim.

Regarding sea-surface temperature (SST), the MAE was improved by the introduction of model cycle 40r1. This cycle included the upgrade to NEMO version 3.4, changes to the vertical mixing in the ocean and the introduction of wave-ocean coupling. However, the MAE increased again with cycle 41r2, together with an increased positive bias, especially in the summer hemisphere. The change was probably an effect of changes in the cloud erosion (see Table A3 above), which had a positive impact of the TOA net solar radiation over sub-tropical oceans but increased the error in the SST. This example illustrates the difficulties in improving a coupled model, when errors are often compensating.

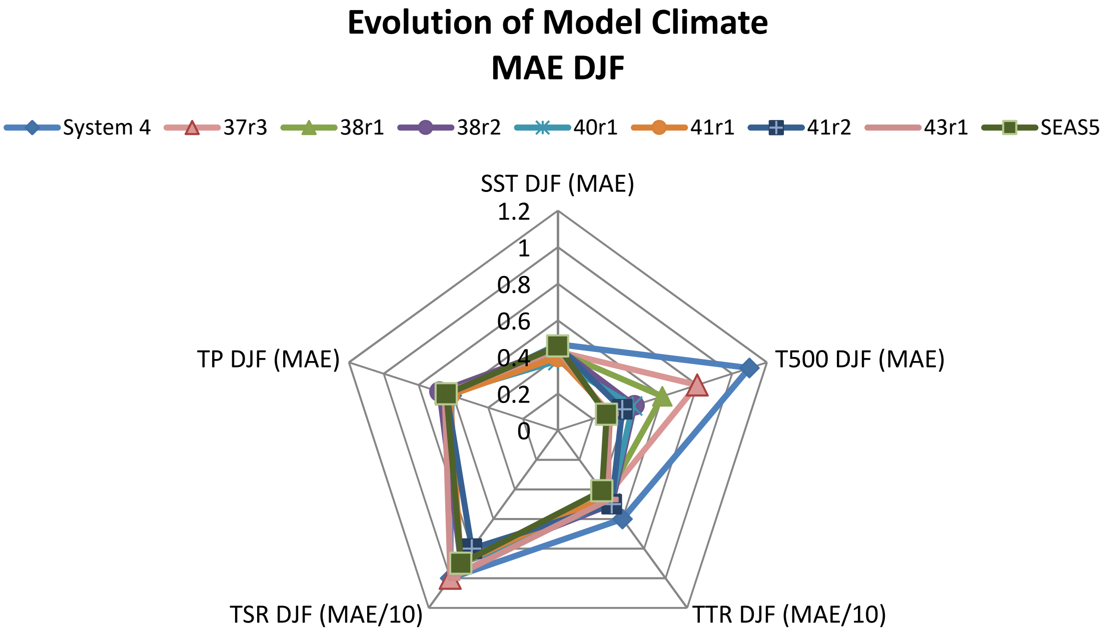
For the TOA net long-wave radiation SEAS5 is better than the TL255 resolution 43r1 over the Maritime continent and western Pacific. The difference is probably related to a better simulation of the Walker circulation with reduced positive precipitation bias over the Maritime continent and better SST bias in the equator cold tongue.

Table A 4 Global mean and mean absolute values of DJF model bias for various fields in seasonal integrations with different IFS cycles and with SEAS5. TTR is top of the atmosphere thermal radiation (or OLR), TSR is top of the atmosphere solar radiation (i.e. the non-reflected part of the incident solar radiation). TP is total precipitation.

Cycle	SST DJF (ME/MAE)	T500 DJF (ME/MAE)	TTR DJF (ME/MAE)	TSR DJF (ME/MAE)	TP DJF (ME/MAE)
S4	-0.03/0.47	-1.1/1.1	4.9/6	0.8/10	0.16/0.61
37r3	0.04/0.43	-0.8/0.8	3.1 /4.5	-1.1/10	0.27/0.64
38r1	0.08/0.43	-0.6/0.6	2.6/4.5	-0.5/9	0.25/0.65
38r2	0.09/0.45	-0.39/0.44	3.6/5	0.7/9	0.23/0.68
40r1	0.07/0.38	-0.35/0.43	3.2/4.9	1.6/9	0.21/0.64
41r1	0.16/0.40	-0.08/0.30	1.6/4.5	0.7/9	0.19/0.62
41r2	0.29/0.46	0.20/0.37	1.4 /5.0	3.5/8	0.22/0.64
43r1	0.16/0.44	0.05/0.29	0.9/4.7	0.27/10	0.22/0.65
SEAS5	0.04/0.46	0.07/0.28	2.0/4.1	0.12/9	0.29/0.64

Table A 5 As for Table A 4 for JJA.

Cycle	SST JJA (ME/MAE)	T500 JJA (ME/MAE)	TTR JJA (ME/MAE)	TSR JJA (ME/MAE)	TP JJA (ME/MAE)
S4	-0.08/0.46	-0.8/0.8	4/5	-1.4/9	0.20/0.66
37r3	0.02/0.44	-0.5/0.6	2.5/4.9	-2.3/9	0.31/0.72
38r1	0.05/0.46	-0.32/0.5	1.9/5	-1.7/9	0.28.0.72
38r2	0.08/0.48	-0.1/0.5	3/5	-0.1/9	0.27/0.73
40r1	0.07/045	-0.01/0.5	3.2/5	1.1/9	0.26/0.72
41r1	0.12/0.44	0.2/0.43	0.9/4.8	-1/8	0.24/0.68
41r2	0.26/0.50	0.5/0.6	0.7/5	1.8/8	0.26/0.70
43r1	0.17/0.48	0.38/0.6	0.29/5	-0.7/9	0.26/0.71
SEAS5	0.00/0.42	0.28/0.44	1.2/4	-2.2/8	0.35//0.74



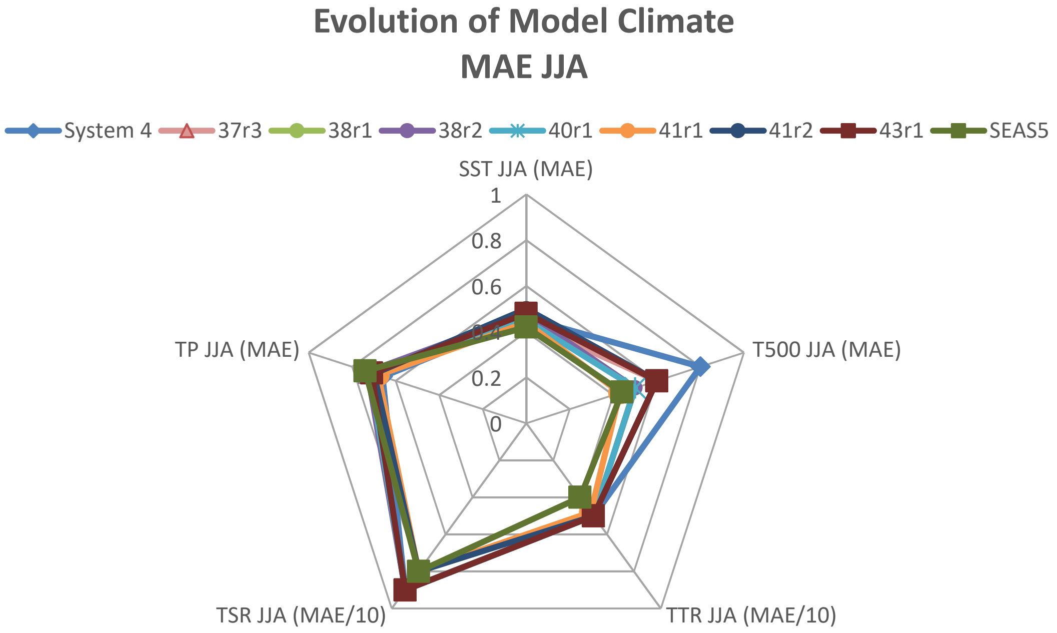


Figure A 9: Graphical representation of the MAE of model climate in DJF (top) and JJA (bottom) corresponding to the MAE values in Table A4 and A5

A.13. MSLP and blocking indices

Figure A 10 compares the MSLP bias in SEAS5 with those of S4. In JJA, high pressure biases in both models (but especially SEAS5) correspond to 500 hPa geopotential height biases in the northern Pacific and Atlantic, suggesting they are related to the displacement of the jet. During DJF, the errors in the amplitude of the stationary planetary waves have considerably reduced in SEAS5. Thus the bias in the North Pacific high disappears almost entirely, and the representation of the winter MSLP trough centred over the British Isles is improved. The improvements in mean MSLP in the North Pacific are accompanied by modest improvements in blocking in the Pacific, as shown using the Tibaldi-Molteni index (Tibaldi and Molteni, 1990) in *Figure A 11*. Atlantic blocking remains unchanged.

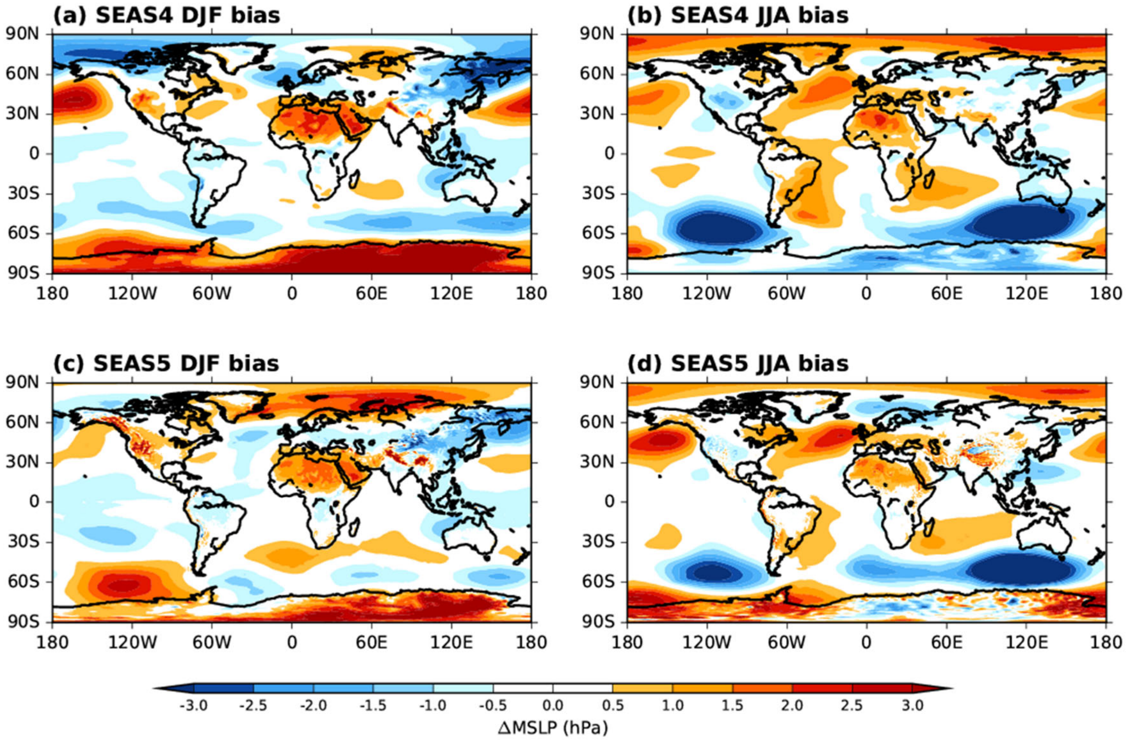


Figure A 10 Winter and summer mean sea level pressure bias in SEAS4 (a,b) and SEAS5 (c,d) with respect to ERAI, for forecasts at 2-4 months. Shown are forecast initialized in November (left) and May (right).

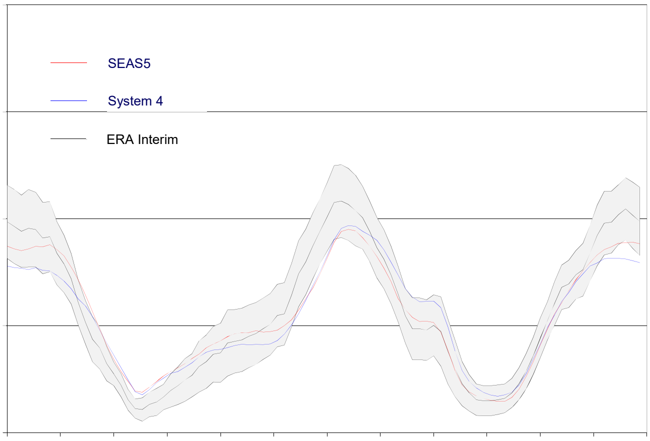


Figure A 11: Daily frequency of blocking in the northern hemisphere based on the Tibaldi-Molteni Index, ERA-Interim in black, System 4 in blue and SEAS5 in red. The grey shading illustrates 95% percent confident intervals for ERA-Interim derived from a Student's t-test.

A.14. Surface Wind and Precipitation biases in the tropics

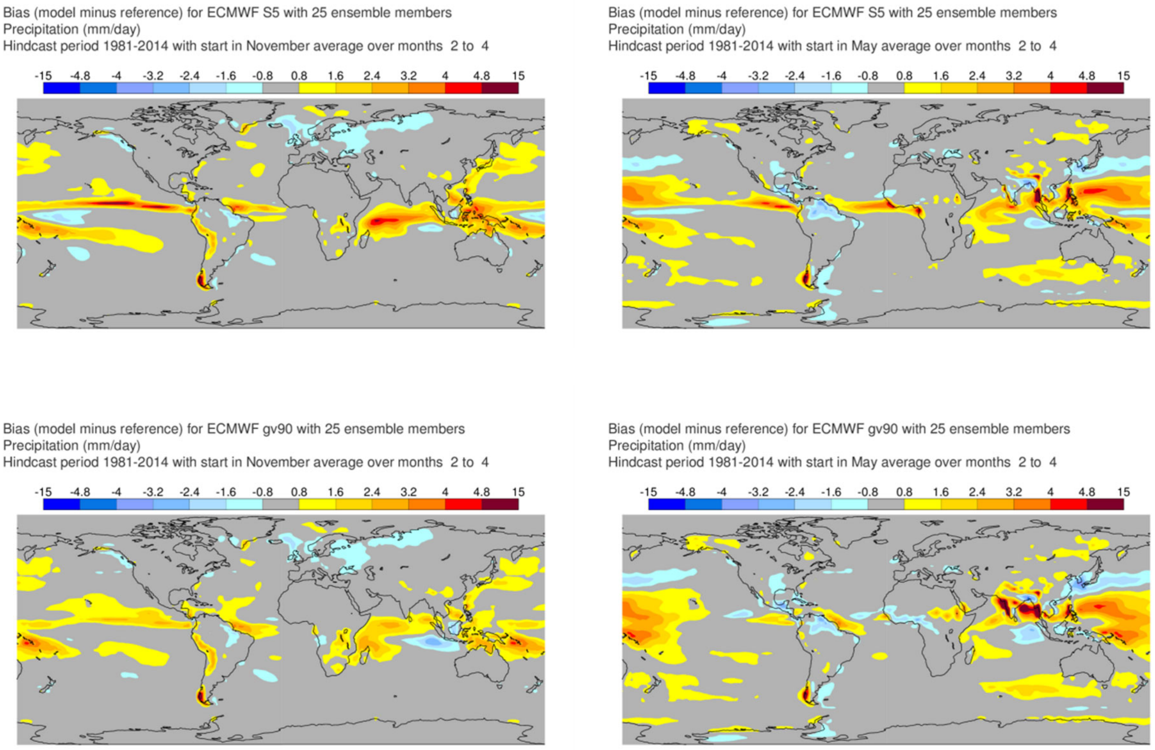

Bias (model minus reference) for ECMWF S5 with 25 ensemble members
Precipitation (mm/day)
Hindcast period 1981-2014 with start in November average over months 2 to 4
Bias (model minus reference) for ECMWF S5 with 25 ensemble members
Precipitation (mm/day)
Hindcast period 1981-2014 with start in May average over months 2 to 4
Bias (model minus reference) for ECMWF gv90 with 25 ensemble members
Precipitation (mm/day)
Hindcast period 1981-2014 with start in November average over months 2 to 4
Bias (model minus reference) for ECMWF gv90 with 25 ensemble members
Precipitation (mm/day)
Hindcast period 1981-2014 with start in May average over months 2 to 4
-15 -4.8 -4 -3.2 -2.4 -1.6 -0.8 0.8 1.6 2.4 3.2 4 4.8 15

Figure A 12 Top) Precipitation biases in SEAS5 for DJF and JJA, for forecast initialized in November and May respectively. The bottom panels show the equivalent biases for experiment SEAS5-ObsSST. The biases are computed with respect the GPCP data set.

During DJF, the SEAS5 surface wind biases over the Indian Ocean are comparatively weaker than in S4 (*Figure 15*, main text). There is a slight westerly bias during DJF, which amounts to an excess of low level convergence over the Maritime Continent. Still, this error is much reduced with respect to S4, where the low-level convergence over the Maritime continent in DJF was even stronger. In contrast, the increased westerly bias along the East African coast in SEAS5 is a degradation with respect to S4.

Over the Equatorial Atlantic and Eastern Pacific, SEAS5 zonal surface wind errors appear as a series of dipoles, displaced northward in the summer season. The errors are consistent with the model circulation being too symmetric around the Equator, lacking meridional asymmetry. This also translates into errors in seasonal migrations of the ITCZ. As in S4, the equatorward component of the trade winds along the eastern coasts is underestimated in SEAS5, especially along the Californian, South American and Benguela coasts, leading to reduced coastal upwelling and warm SST biases. These wind biases are also present in the uncoupled model (not shown).

Around the Maritime Continent and warm pool region, biases in the wind are associated with precipitation biases. As for S4, SEAS5 shows a pronounced dry bias over the Equatorial Pacific west of the dateline, related with the easterly bias and westward extension of the cold tongue. Over the ITZC regions in all ocean basins there is instead an excess of precipitation in the coupled model. These two errors are related to the coupling, not being present in uncoupled model (see Figure A 12). Over the Maritime continent the precipitation in SEAS5 is more realistic than in S4 (not shown), in both position and intensity, although it appears slightly skewed towards the Western Pacific, where there is an excess of precipitation at expense of a slight dry bias in the Eastern Indian Ocean (IND2), which is present in the uncoupled experiment. As discussed above, during JJA this dry bias in SEAS5 is associated with an easterly bias and a cold SST. Over the broad Indian Ocean there is an excess of precipitation in SEAS5, to which the warm SST biases probably contribute. In JJA there is also an excess of precipitation north off the Equator in the Western Pacific, associated with the location of the ITCZ and monsoon circulations; the latter is even stronger in the uncoupled integrations.

A.15. Impact of stochastic physics in the tropics

For forecasts initialized in May, switching off stochastic physics leads to a cooling of the Niño 3.4 region in all forecast configurations. This is also true for the November starts, but in this case the impact of stochastic physics is weaker. When we look at the r.m.s errors of NINO3.4 SST forecasts, switching off stochastic physics leads to a substantial rise in r.m.s.e., as well as a reduction in ensemble spread. The stochastic physics is beneficial for the mean state, the accuracy of the ENSO forecasts and the forecast reliability.

In line with previous findings (Weisheimer et al, 2014), stochastic physics acts to increase substantially the spread of the MJO in the first few weeks of the forecast, leading to a more realistic growth of uncertainty in the coupled system, as shown in *Figure A 13*.

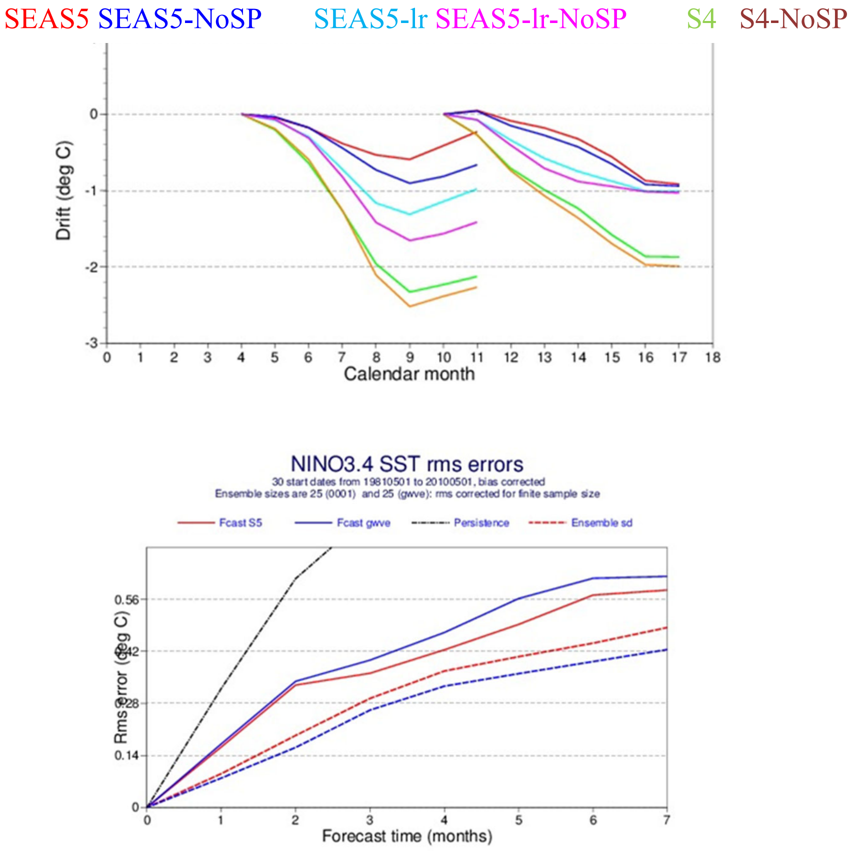


Figure A 13: a) Impact of stochastic perturbations on SST drift in Nino3.4 for May and November start dates. SEAS5 (with and without stochastic physics) are shown in red and blue respectively; S4 and its non-stochastic physics equivalent are shown in green and orange; the drift for the low-resolution configuration of SEAS5, with and without stochastic, is indicated by the cyan and purple lines. b) Evolution of RMSE (solid) and ensemble spread (dashed) of Nino3.4 SST over lead time for May start dates showing SEAS5 in red and SEAS5 without stochastic physics perturbations in blue.

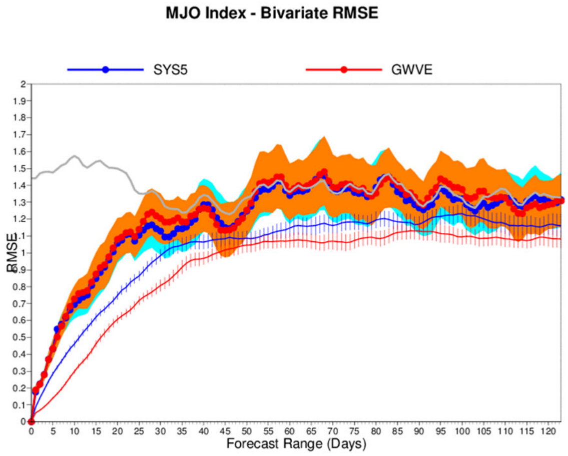


Figure A 14: Impact of stochastic perturbations on the evolution of the RMSE (two top lines with uncertainty shadings) and ensemble spread (two bottom lines with error bars) of the bivariate MJO index. SEAS5 in blue, SEAS5_noSP in red

A.16. Ocean resolution

To allow an evaluation of the impact of changes in vertical and horizontal ocean model resolution between SEAS5 and S4 separately from the question of initial conditions, we generate balanced initial conditions for the ORCA1_Z42 and ORCA1_Z75 NEMO configurations (see *Table A 6*) that are consistent with the ORAS5 reanalysis (*Table A 7*). This is done by running additional NEMO experiments at the target resolution with the same surface forcings as ORAS5 and a 5-day relaxation of 3D ocean temperature and salinity, sea-ice thickness, and sea-ice concentration towards interpolated monthly mean values from ORAS5. The resulting experiments provide balanced ocean state estimates suitable for initializing ORCA1 configurations that faithfully reproduce the variability of surface and sub-surface ocean properties present in ORAS5. These low-resolution analogues of ORAS5 are then used to initialize re-forecasts with the SEAS5 atmospheric configuration coupled to the NEMO ORCA1_Z42 and ORCA1_Z75 configurations (*Table A 8* and *Table 4*).

SST biases from S4, SEAS5, and the SEAS5.ORCA1 re-forecasts are shown in Figure A 15. SST biases in the equatorial Pacific and north-west Atlantic are substantially improved in SEAS5 compared to S4. These improvements can be attributed to a combination of increased ocean horizontal resolution and improvements to the atmospheric model. The increase in ocean vertical resolution from 42 to 75 levels has very little impact on SST biases at seasonal lead times.

Table A 6 NEMO ocean model configurations

	ORCA1_Z42	ORCA1_Z75	ORCA025_Z75
Nominal resolution (ni x nj)	100 km (362 x 292)	100 km (362 x 292)	25 km (1442 x 1021)
Vertical levels (thickness of top level)	42 (10 m)	75 (1 m)	75 (1 m)
NEMO time step	60 min	60 min	20 min

Table A 7 Nemo experiments used to initialize ocean in SEAS5 re-forecasts

	gro6	gro7	ORAS5
NEMO configuration	ORCA1_Z42	ORCA1_Z75	ORCA025_Z75
Forcing	ERA/OPS + wave	ERA/OPS + wave	ERA/OPS + wave
Assimilation	5-day nudging towards ORAS5 monthly mean T/S/ice + SST restoration	5-day nudging towards ORAS5 monthly mean T/S/ice + SST restoration	NEMOVAR + SST restoration + bias correction.
Period	1979-2015	1979-2015	1979-present
Ensemble	5 members	5 members	5 members

Table A 8 Seasonal re-forecasts with SEAS5 combine with different ocean model resolutions

	SEAS5.ORCA1_Z42 (guy1)	SEAS5.ORCA1_Z75 (guuk)	SEAS5
Atmosphere	Tco319 L91, tstep=1200	Tco319 L91, tstep=1200	Tco319 L91, tstep=1200
Ocean	NEMO ORCA1_Z42, tstep=3600	NEMO ORCA1_Z75, tstep=3600	NEMO ORCA025_Z75, tstep=1200

Ocean initial conditions	gro6	gro7	ORAS5/OCEAN5
Start dates, length	May/Nov, 7 months	May/Nov, 7 months	Monthly, 7 (13) months
Period	1981-2016	1981-2016	1981-present
Ensemble	5 members	5 members	25 members

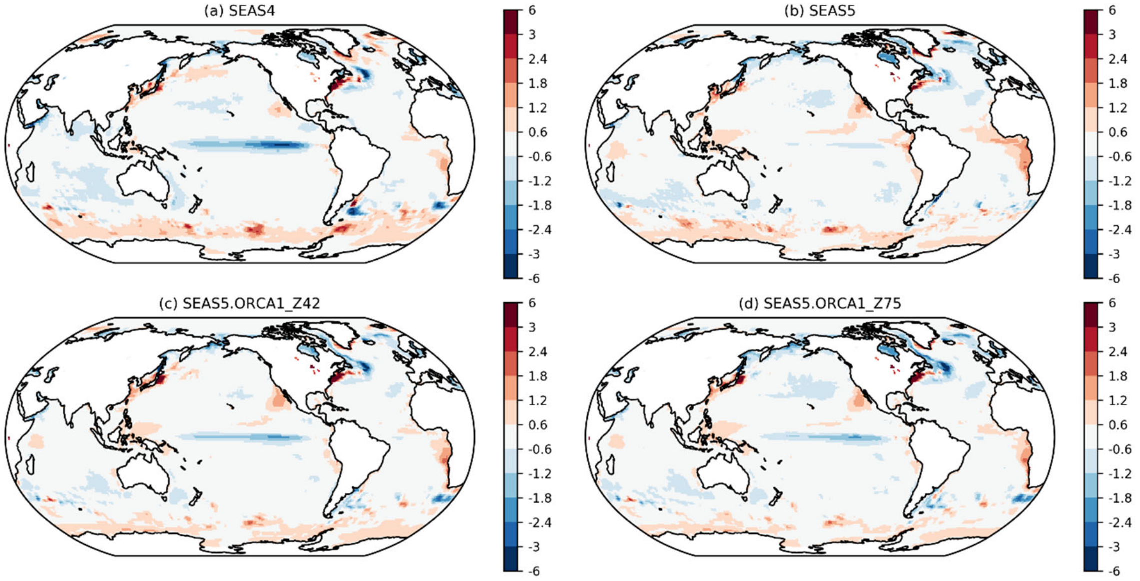


Figure A 15 SST biases (K) relative to ERA-interim calculated using forecast month 4 from November and May starts averaged together, over the period 2000-2015. Differences between S4 and SEAS5 SST biases can be attributed to a combination of increased ocean horizontal resolution and improvements to the atmospheric model. The increase in ocean vertical resolution from 42 to 75 levels has very little impact on SST biases at seasonal lead times.

In the North Atlantic, the reduced SST biases in SEAS5 are a consequence of increased horizontal resolution in the ocean and associated improvements to the large-scale overturning circulation and position of the Gulf Stream (Figure A17a). These changes to SST are accompanied by local responses in 2m air-temperature, turbulent heat fluxes, precipitation, and mean sea level pressure (Figure A 16b-e). This is consistent with previous studies reporting the benefits of eddy-permitting ocean model resolutions for improved representation of ocean boundary currents, mesoscale eddies, air-sea interaction, topographically controlled flows (e.g. Bryan et al, 2010; Hewitt et al, 2016; Roberts et al, 2016) in multi-decadal climate integrations. However, it is less clear whether these changes influence the atmospheric variability or if there are remote impacts on the mean climate of Europe at seasonal time-scales. For example, Scaife et al (2011) found that increased ocean model resolution and improved SST biases in the Met Office climate model were associated with greatly improved Atlantic winter blocking frequency in the Euro-Atlantic sector. We have found a similar response in multi-decadal integrations with the same version of the IFS coupled model performed as part of the PRIMAVERA project (Figure A 16f; Roberts et al, 2018). However, winter blocking frequencies in seasonal experiments do not show this sensitivity with both SEAS5 and SEAS5.ORCA1_Z75 exhibiting blocking frequencies similar to those in atmosphere-only experiments forced by observed SSTs (Figure A 16f).

This apparent contradiction can be explained by differences in the magnitude, and possibly the spatial pattern, of SST biases in seasonal and multi-decadal integrations and serves as an example where impacts of increased ocean resolution may be lead-time dependent.

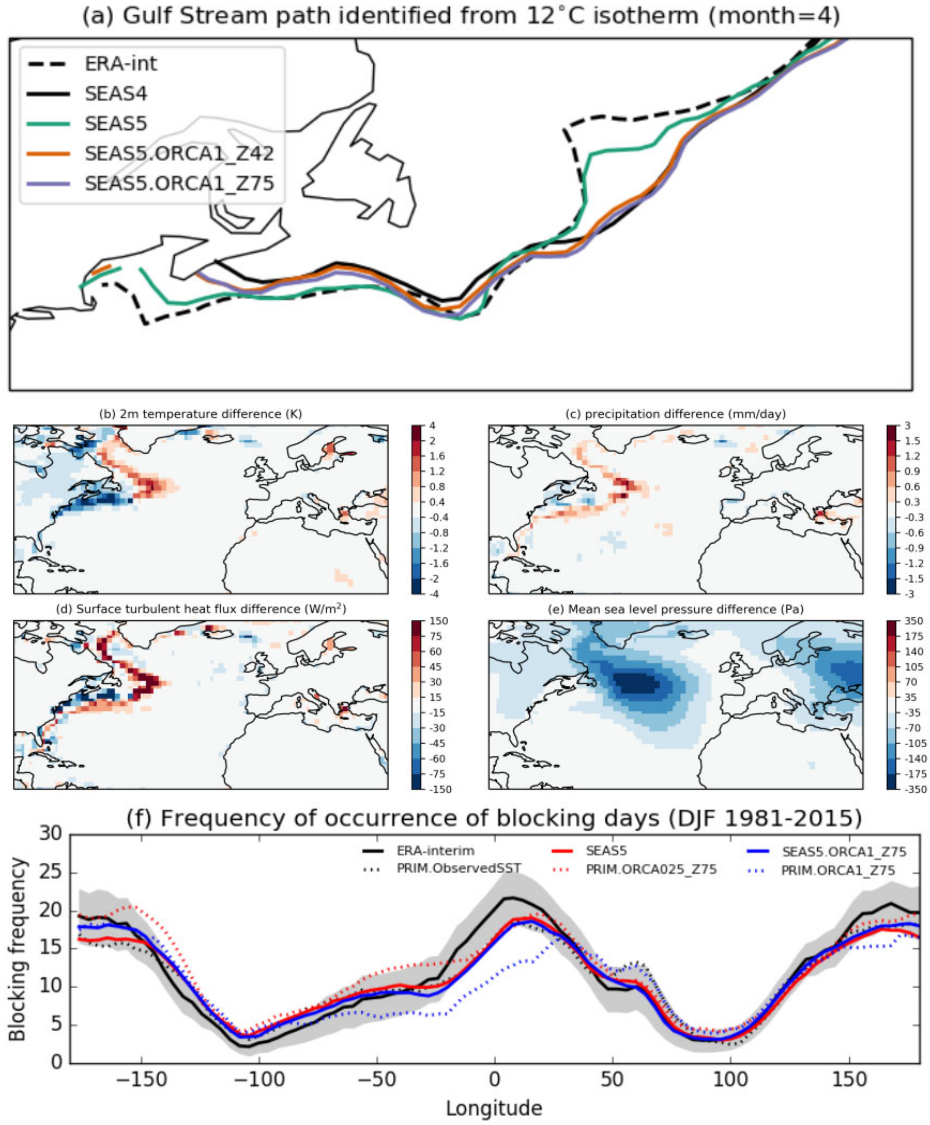


Figure A 16 (a) Position of Gulf Stream/North Atlantic Current identified from the 12°C isotherm in month 4 sea surface temperature forecasts . (b-e) Impact of ocean resolution on atmospheric fields calculated in month 4 (SEAS5 minus SEAS5.ORCA1_Z75) for November starts over the period 1981-2015. Note that although the absolute value of biases is sensitive to the specified period, the difference between SEAS5 and SEAS5.ORCA1_Z75 is relatively stable. (f) Frequency of winter blocking days estimated using the Tibaldi and Molteni (1990) index. Dotted lines correspond to multi-decadal experiments performed as part of the PRIMAVERA project using the same cycle of the IFS (CY43R1) but with reduced atmospheric resolution (Tco199) compared to SEAS5 (see Roberts et al, 2018 for further details).

The dependence of biases on lead-time means that seasonal forecast bias is not a reliable indicator of the asymptotic behaviour of the coupled model, and that longer integrations are needed to assess the impact and quality of an ocean model within a coupled Earth system model. The asymptotic behaviour of coupled models is becoming more relevant to ECMWF with the development of coupled approaches to Earth system reanalysis. The quality of the coupled model is critical in such reanalyses because of its influence as a background field, even more so during periods and/or regions with limited observational constraints.

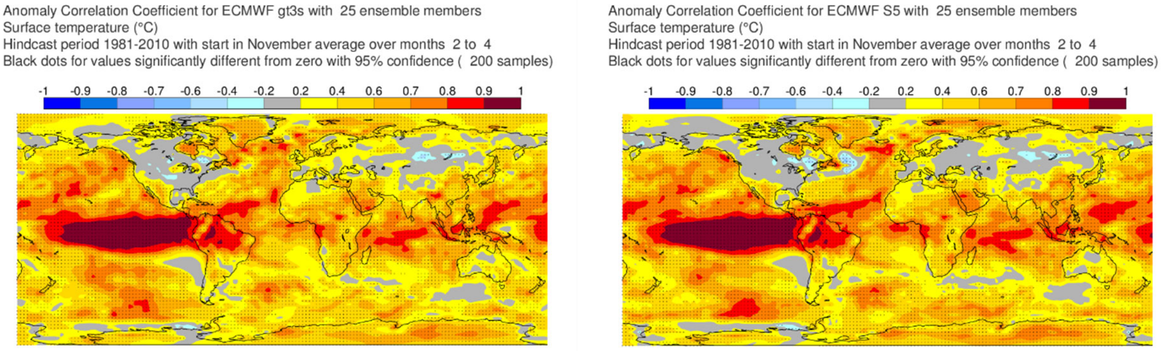

Anomaly Correlation Coefficient for ECMWF gt3s with 25 ensemble members
Surface temperature (°C)
Hindcast period 1981-2010 with start in November average over months 2 to 4
Black dots for values significantly different from zero with 95% confidence (200 samples)
-1 -0.9 -0.8 -0.7 -0.6 -0.4 -0.2 0.2 0.4 0.6 0.7 0.8 0.9 1
Anomaly Correlation Coefficient for ECMWF S5 with 25 ensemble members
Surface temperature (°C)
Hindcast period 1981-2010 with start in November average over months 2 to 4
Black dots for values significantly different from zero with 95% confidence (200 samples)
-1 -0.9 -0.8 -0.7 -0.6 -0.4 -0.2 0.2 0.4 0.6 0.7 0.8 0.9 1

Figure A 17: T2m anomaly correlation skill for DJF with SEAS5-lr (left) and SEAS5 (right), showing differences in the NW Atlantic and over Europe.

A.17. Estimates of AMOC variability

It is difficult to ascertain the realism of the AMOC strength for the full reforecast period, because it has been observed only since 2005 with the installation of the RAPID mooring array. However, there have been measurements of the Gulf Stream strength at the Florida Strait since the 1980s by virtue of telecommunications cables, so a direct comparison of modelled and observed current strengths for the early period is possible at that location. The Gulf stream is a strong contributor to the AMOC, transporting warm and saline water from the Gulf of Mexico into the North Atlantic. Figure A18 presents time series of AMOC strength and Florida Strait transports in the ocean simulations that provide the initial conditions for S4, SEAS5, as well as the two experimental reforecast sets Ctrl-SST and Ctrl-noSST discussed previously. The AMOC since 2005 is very well represented by ORAS5, whereas Ctrl-noSST underestimates it. In the early period, the AMOC in ORAS5 was considerably stronger, but there are no observations to verify the realism of that. However, there is good coherence in ORAS5 between the time series of AMOC and the Florida Strait transport, and for the latter there are observations available going back to the 1980s. In comparison to these observations, ORAS5 appears to overestimate the Florida Strait transport in the 1980s, which suggests that the AMOC may also be overestimated. However, the Florida Strait transport in Ctrl-noSST has the same temporal variability as ORAS5, in spite of not having assimilation nor relaxation to SST (Figure A18b), being driven only by winds and surface fluxes. Also, the observational record has a discontinuity at the time of the modelled changes, and it is possible that calibration issues may play some role in the discrepancies. More investigation is needed before final conclusions can be drawn. Nonetheless, the connection between Gulf Stream intensity and AMOC is stronger in ORAS5 than in the control model integrations.

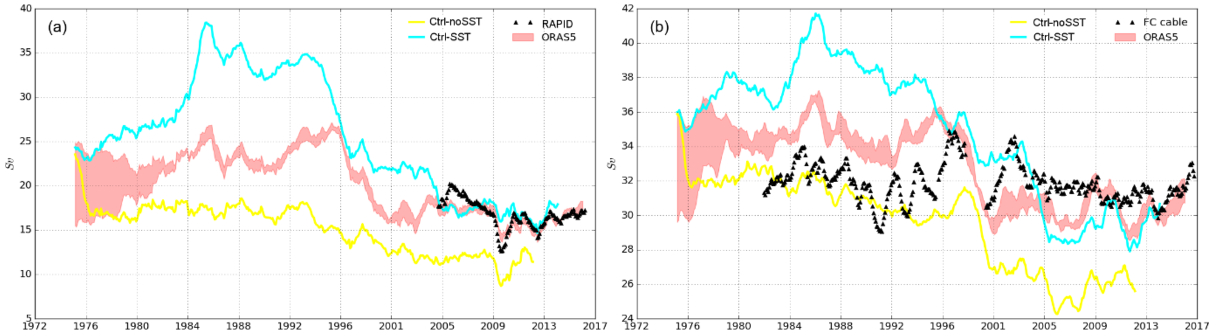


Figure A 18: Ocean volume transport (a) in the upper 1000m meters of the Atlantic at 26N (AMOC), (b) through the Florida Strait in ORAS5 and the two Control-experiments with and without SST nudging. Black triangles show available observations from (a) the RAPID array of moorings and (b) the Florida Strait cable. ORAS5 is shown in a with a range to indicate the minimum and maximum of its 5 ensemble members.

A.18. CCA of co-variability of North Atlantic SST and precipitation

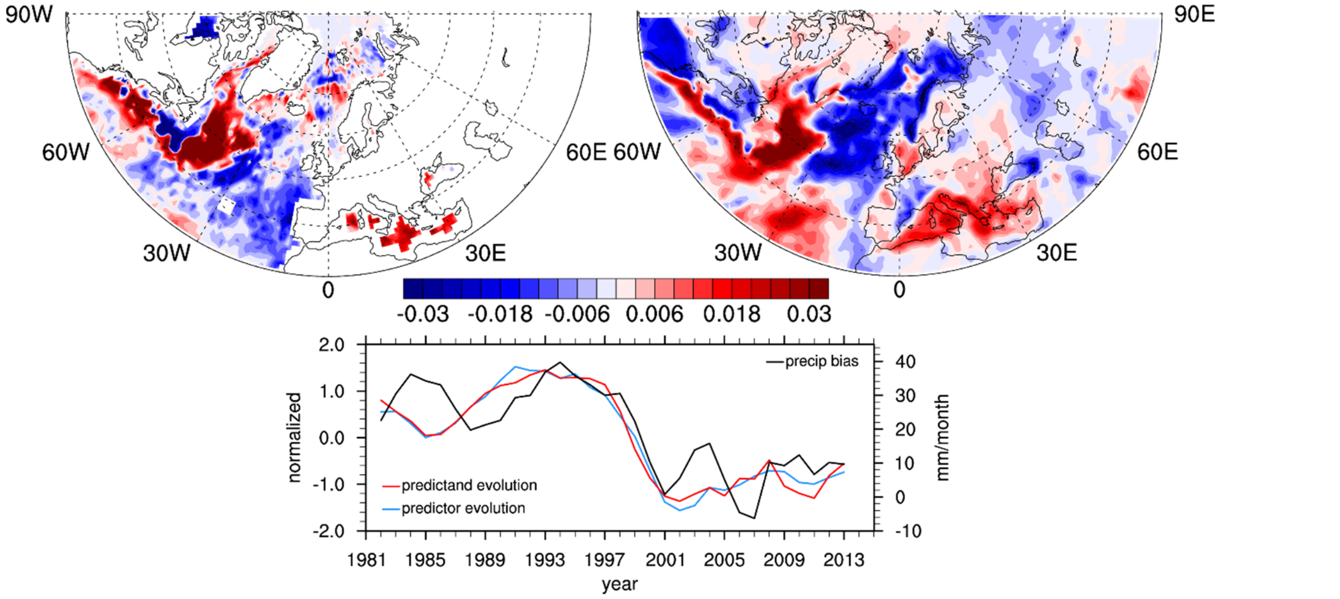
<table>
<tr><th>Predictor: SEAS5 Nov SST error (27% of total variance explained)</th><th>Predictand: SEAS5 DJF precipitation error (18% of total variance explained)</th></tr>
<tr><td colspan="2">
</td></tr>
</table>

Figure A 19 First canonical patterns (normalized) and associated time series (r=0.99) for SEAS5 SST error in November and precipitation error in DJF for all November starts 1981-2014.

Canonical Correlation Analysis (CCA, e.g. Wilks 2011) objectively selects coherent modes of temporal co-variability and measures the percentage of variance. CCA has been applied to the relation between SEAS5 SST and precipitation errors in the NH extra-tropics in the region 90W-90E. The dominant CCA mode closely follows the time evolution of the NASD-averaged SEAS5 precipitation error - which exhibits a decrease of about 20-30mm/month between 1995-2000- and explains about 18% of the variance over the half-hemisphere. This mode's spatial pattern is clearly dominated by SST errors in the Subpolar Gyre, with pronounced precipitation error co-located with SST (Figure A 19). This is consistent with the change in SST errors driving large-scale changes in precipitation errors across the Atlantic basin.

A.19. QBO Teleconnections to the NH

Observations suggest a fairly strong relationship between the QBO and the NH winter polar vortex (Holton-Tan effect), and also a connection to the NH surface circulation. We study these using a composite of QBOE-QBOW years, defined from observations using a multi-level phase angle to define the QBO phase (Hamilton and Boer, 2008).

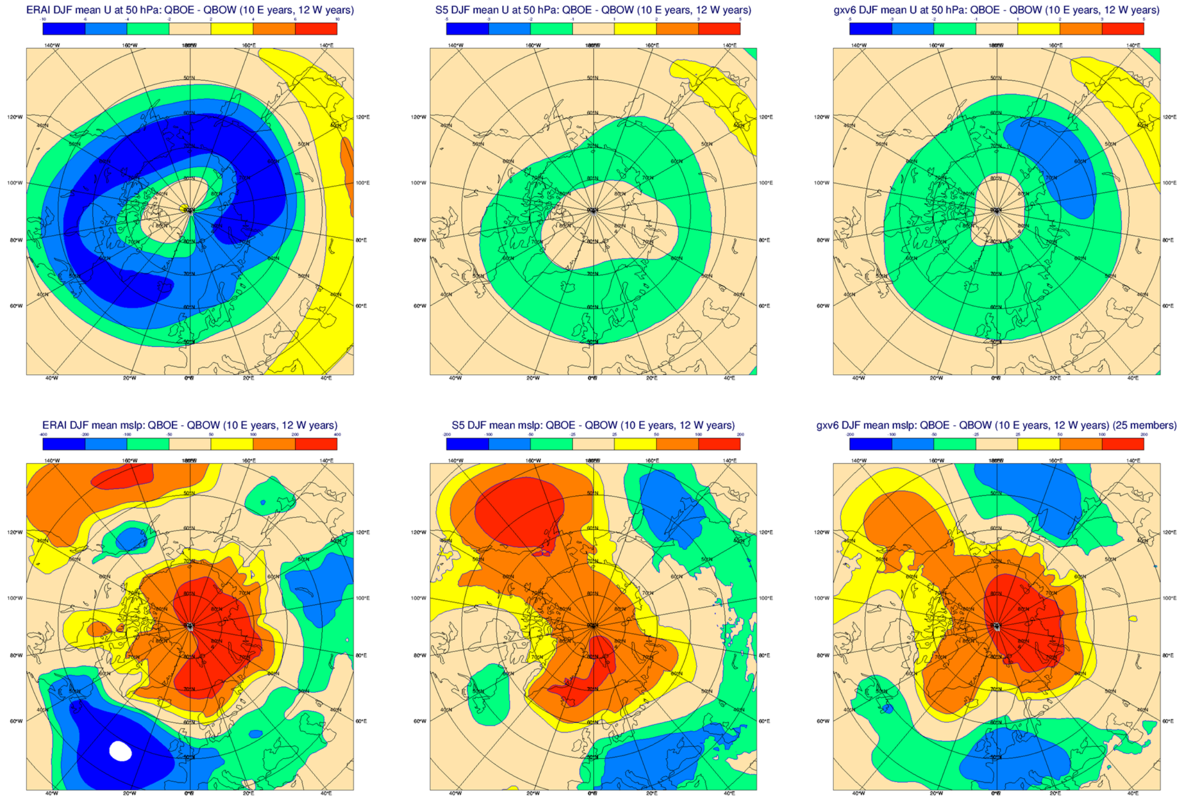

ERAI DJF mean U at 50 hPa: QBOE - QBOW (10 E years, 12 W years)
S5 DJF mean U at 50 hPa: QBOE - QBOW (10 E years, 12 W years)
gxv6 DJF mean U at 50 hPa: QBOE - QBOW (10 E years, 12 W years)
ERAI DJF mean mslp: QBOE - QBOW (10 E years, 12 W years)
S5 DJF mean mslp: QBOE - QBOW (10 E years, 12 W years)
gxv6 DJF mean mslp: QBOE - QBOW (10 E years, 12 W years) (25 members)

Figure A 20: TOP) QBOE-QBOW composite for NH winter zonal winds at 50 hPa from ERAI (left), SEAS5 (centre) and SEAS5 L137 (right). Note that contour interval for the SEAS5 models is half that for ERAI. BOTTOM) QBOE-QBOW composite for NH winter MSLP from ERAI (left), SEAS5 (centre) and SEAS5 L137 (right). Again, contour interval for the SEAS models is half that for ERAI.

As shown in Figure A20, SEAS5 reproduces the Holton-Tan effect in the stratospheric vortex, but with a substantially weakened amplitude, only about 1/5 of that observed. By contrast, S4, which also reproduced the connection, did so with an amplitude about 1/3 of that observed (not shown). The surface impact over the Arctic is also reproduced, again too weak, but interestingly not as weak as the stratosphere response itself. This suggests that part of the influence of the QBO on the surface may not be via the stratospheric vortex, but a different pathway. The observed response includes strong negative values over the Atlantic (up to 4 hPa) which is not seen in the model.

Results from a L137 version of SEAS5 have a stronger response in the stratosphere than SEAS5, and a slightly stronger and better structured response at the surface (right hand panels of Figure A20). It should be noted that although this L137 experiment has improved QBO structure at the 10 hPa level (due to the extra resolution), the QBO in the lower stratosphere is still strongly damped by vertical diffusion, which may be affecting the strength of the QBO teleconnections. Further work is needed to understand the pathways from the QBO to the troposphere. Improvements to the vertical structure of the QBO, if needed, are expected to require a re-formulation of the vertical diffusion.