Technical Memo

ECMWF
European Centre for Medium-Range Weather Forecasts

854

Extended-range Prediction

Frédéric Vitart, Magdalena Balmaseda, Laura Ferranti, Angela Benedetti, Beena Sarojini, Steffen Tietsche, Junchen Yao, Martin Janousek, Gianpaolo Balsamo, Peter Bechtold, Martin Leutbecher, Inna Polichtchouk, David Richardson, Christopher Roberts and Tim Stockdale (Research Department)

November 2019

Series: ECMWF Technical Memoranda

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

Contact: library@ecmwf.int

© Copyright 2019

European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, RG2 9AX, UK

Literary and scientific copyrights belong to ECMWF and are reserved in all countries. This publication is not to be reprinted or translated in whole or in part without the written permission of the Director-General. Appropriate non-commercial use will normally be granted under the condition that reference is made to ECMWF.

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

Abstract

The production of extended-range forecasts is now a consolidated operational activity at ECMWF. It bridges the medium and the seasonal range from the perspective of users, model development and initialization options. This paper reviews five years of progress in the ECMWF forecasts at this time range and provides a summary of main lessons learns from the scientific exploitation of the S2S data base. A proposal for the extended-range forecasting system on the new HPC is presented. Developments for a longer time frame are described. They are being explored in the context of the second phase of the WWRP/WCRP sub-seasonal to seasonal (S2S) prediction project, which provides an opportunity to coordinate this research with the international community through coordinated experiments and analysis of the S2S database.

1 Introduction

Extended-range forecasts have been produced routinely at ECMWF since March 2002, and operationally since October 2004. Since 2002, the configuration of the extended-range forecasting system has changed several times. In the current configuration, the extended-range forecasts are generated by extending the 15-day ensemble integrations to 46 days twice a week (at 00 UTC on Mondays and Thursdays) (See Table 1). This extended-range is also known as sub-seasonal range. Forecasts for the medium-range and extended-range are integrated in the ECMWF ensemble forecasting system (ENS). ENS includes 51 members run with a horizontal resolution of Tco639 (about 18 km) up to forecast day 15 (referred to as LegA), and Tco319 (about 36 km) thereafter (referred to as LegB). Initial perturbations are generated using a combination of singular vectors, and perturbations generated using the ECMWF ensemble of data assimilations. A stochastic physics scheme (SPPT) is used to represent model uncertainty (Lock et al, 2019).

The last report to the ECMWF science Advisory Committee (SAC) on extended-range prediction (available as ECMWF Technical Memorandum 738[1]) was produced in 2014. This report showed that the skill of the sub-seasonal forecasts at ECMWF had improved significantly since the first operational extended-range forecasts in 2004. This improvement could be linked to improved skill to predict the Madden Julian Oscillation (MJO), an important source of predictability for the sub-seasonal time range with an average gain of about 1 day of prediction skill per year and improved tropical-extra-tropical teleconnections associated to the MJO. The skill of the ECMWF monthly forecasts to predict the North Atlantic Oscillation (NAO) had also increased over the years, albeit at slower pace. The skill for predicting sudden stratospheric warmings (SSW), another source of sub-seasonal predictability, had also improved over the previous 10 years, although the downward propagation associated with SSWs was much weaker in the model than in ERA Interim. Extended-range forecasts displayed some skill in predicting heat waves, although with a weaker amplitude and mostly when the anti-cyclonic circulation was already present in the initial conditions. The report argued that improvements in the ECMWF sub-seasonal forecasts were expected with the planned introduction of a dynamic sea-ice model, increased atmospheric and oceanic resolutions and the extension of the re-forecasts (twice a week and 11 members compared to once a week with 5 members previously).

Since 2014, the planned changes mentioned above have been implemented, together with new versions of the atmospheric model. In addition, extended-range forecasting has now a much higher profile in the ECMWF strategy than in 2014 and new headline scores based on extended-range prediction have been implemented. The S2S database opened in May 2015 and has enabled comparisons with extended-range forecasts from other operational centres. Besides, the implementation of the next high-performance

[1] https://www.ecmwf.int/sites/default/files/elibrary/2014/12943-sub-seasonal-predictions.pdf

computer (HPC) in Bologna provides an opportunity to revise our ensemble strategy for extended-range prediction. Forward looking plans should be seen in the context of enhanced international efforts in this area (e.g. WWRP/WCRP S2S project moving to phase 2). All these factors make it timely to revisit the current status and future prospects of extended-range prediction at ECMWF.

This memorandum is organized as follows: Section 2 describes the main changes in the forecasting system since 2014 and compares the performance of the current system with the system which was operational five years ago. Section 3 shows how the extended-range forecasts at ECMWF compare with those from other operational centres (S2S Database). Section 4 presents a few case studies of extended-range forecasts. Section 5 discusses some important issues affecting the forecast skill at this time range and Section 6 presents future plans. The main results of this report are summarized in Section 7.

Cycle **Implementation Date**	**CY40R1** **19/11/2013**	**CY41R1** **14/05/2015**	**CY41R2** **08/03/2016**	**CY43R1** **22/11/2016**	**CY43R3** **11/07/2017**	**CY45R1** **06/06/2018**	**CY46R1** **11/06/2019**
Time Range	32 days	46 days	-	-	-	-	-
Atmospheric Resolution	Leg A: up to day 10 Tl639L91 Leg B: After day 10 Tl319L91	-	Leg A: up to day 15 Tco639L91 Leg B: up to day 46 Tco319L91	-	-	-	-
Ocean Resolution	1 degree 42 levels	-	-	¼ degree 75 levels	-	-	-
Prognostic Sea-Ice	No	No	No	Yes	-	-	-
Forecast Frequency	Thursdays and Mondays	-	-	-	-	-	-
Re-forecast Frequency	Thursdays	Thursdays and Mondays	-	-	-	-	-
Re-forecast ensemble size	5	11	-	-	-	-	-
Reforecast initialization	ERA-Interim ERAI-Land ORAS4	- - -	- - -	- - ORAS5	- - -	- - -	ERA5 ERA5 -

Table 1: Summary of the main changes in the configuration of the extended-range real-time forecasts and re-forecasts since Nov 2013.

2 Evolution Extended-range forecasts since 2014

Several changes to the model physics have been applied to the operational extended-range forecasts since 2014 as described in Table 1. Since 2014, several changes to the model physics changes have been introduced. A new and faster radiation scheme "ecRad" (Hogan and Bozzo, 2018, implemented in cycle 43R3) includes longwave scattering and is called with a frequency of 1 hour. The use of approximate updates to the radiative fluxes every model timestep and for every gridpoint leads to improved 2-m temperatures at coastlines. A new fully 3D aerosol climatology from CAMS is now used (cycle 46R1), including improved optical properties. The new ozone climatology, also from CAMS, together with other radiation improvements, have reduced the upper stratosphere warm bias by around 5K (Hogan and Bozzo, 2018) and the lower stratospheric cold bias by approximately 0.5 K (cycle 43R3). Further improvements to stratospheric temperature and winds have been achieved by a reduction of the amplitude of non-orographic waves at launch-level (level from which the waves are initialized) for the 91-level model version. Stratospheric water vapour has been improved (increased) in the middle to upper stratosphere taking account of the higher amount of methane now observed in the stratosphere. The numerics and formulation of the warm rain microphysics have been thoroughly revised (Ahlgrimm and Forbes, 2014; Forbes, 2018) improving the overall distribution of surface rain rates and producing a more continuous and realistic inland extension of rain (cycle 45R1). Furthermore, the spurious excessive low cloud cover in Arctic regions has been reduced. Revisions to the convection scheme include a more realistic formulation of mixed-phase clouds and the glaciation and melting of rain, significantly improving (warming) the upper troposphere, thereby also improving the upper-tropospheric winds (cycle 43R3). Finally, the near surface temperature biases have also been reduced by retuning the soil thermal conductivity.

The impacts on the extended-range forecasts of each model cycle as well as individual changes to the ensemble forecast configuration are performed routinely and discussed in the final documentation on each operational cycle[2]. The testing of the operational model cycles, based on a large set of 5-member ensemble re-forecasts starting once a month and covering 25 years (see Buizza et al, 2018 for more details), indicated in general neutral to slightly positive trends in the extended-time range forecast skill scores. The incremental changes are rarely statistically significant except for cycle 45R1 which shows a statistically significant improvement in week 1 for tropical winds and positive effect on skill across all parameters in the European domain. The MJO Index ensemble was significantly under-spread, but changes in 45R1 to the SPPT scheme have brought close agreement between spread and error throughout the 46-day forecast range (see Lock et al, 2019 for more details). The underestimation of the MJO Index amplitude error has been significantly improved throughout the forecast.

Although the impact of a new model cycle on the extended-range forecast skill scores is generally small, the cumulated impact of model cycles over several years is likely to be more visible. In order to evaluate the impact of the various changes to the extended-range forecasting system since 2014, the operational re-forecasts produced between November 2018 and 2019 (cycle 45R1) have been compared with reforecasts produced 5 years earlier from November 2013 to March 2014 (cycle 40R1). The use of re-forecasts to assess the evolution of the extended-range forecast performances, instead of real-time forecasts, is justified by the fact that the skill of the extended-range forecasts displays a very strong interannual variability which is often much larger than the difference of skill we want to measure. Therefore, re-forecasts of the operational model in 2013/2014 and 2018/2019 have been compared over

[2] https://confluence.ecmwf.int/display/FCST/Changes+to+the+forecasting+system

the common 1999-2012 period. To make the comparison fair, the same re-forecast frequency (once a week) and ensemble size (5 members) were used for this evaluation. The period November to March has been chosen so that it includes a single model cycle (cycle 40R1 was implemented in November 2013).

2.1 Biases

The model biases have changed significantly since 2014. The cold temperature biases in the Tropics in cycle 40R1 (2014 integrations) have been significantly reduced (Figure 1). The same holds for the very strong positive bias in the upper stratosphere. However, the current system displays a stronger cold bias at 50 hPa, which could be a consequence of the increased horizontal resolution in 2016 (see Polichtchouk et al, 2019). The biases over the South Pole have also been slightly reduced. In addition, the current system displays smaller zonal wind biases in the Tropics in the upper troposphere (not shown). These changes in model climate, which are particularly strong in the Tropics, are likely to have an impact on the extended-range forecast skill scores.

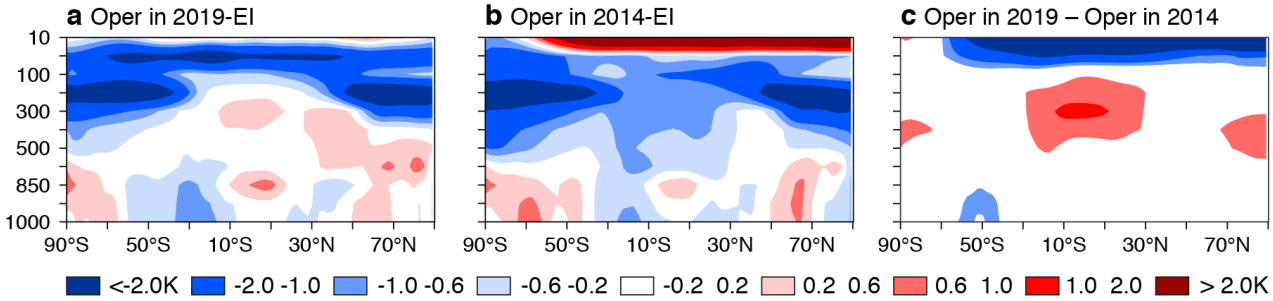

Figure 1: Longitude-averaged temperature biases at lead time of 26-32 days as a function of pressure (y-axis) and latitude (x-axis) relative to ERA Interim of reforecasts produced between November 2018 and March 2019 (left panel) and between November 2013 and March 2014 (middle panel). The right panel shows the difference when it is statistically significant using a 10,000 bootstrap re-sampling procedure. The re-forecast period is 1999-2012.

2.2 Extended-range forecast skill scores

2.2.1 Headline scores

Because of the low signal-to-noise ratio of real-time forecast verification in the extended-range, re-forecasts are a useful additional resource for documenting trends in skill. The top panel of Figure 2 shows the skill of the ENS in predicting 2m temperature anomalies in week 3 in the Northern Extratropics. This is an additional headline score of ECMWF which was recommended by the Technical Advisory Committee (TAC) Subgroup on Verification. The plot shows the discrete ranked probability skill score (Weigel et al, 2007) of the weekly mean of the 2-metre temperature of the ECMWF reforecasts verified against ERA-Interim analyses (the red line) and SYNOP observations (the magenta line). The skill against observations is the ECMWF additional headline score.

Verification against both observations and analyses shows that there has been a substantial increase in skill from 2005 to 2012, and little change (against analysis), and a slight decrease (against observations) thereafter. Note that the verification is based on a sliding 20-year period and is therefore less sensitive to changes from year to year than the real-time forecast evaluation, but some sensitivity remains, e.g. due to major El-Niño events falling within, or dropping out of, the sliding period.

In addition to 2-metre temperature, ECMWF also routinely verifies the extended-range forecast of precipitation and 10m wind speed. The middle and lower panels of Figure 2 show that the level of skill

for these two parameters is lower than for 2m temperature, and just marginally positive. However, there is a positive trend in the precipitation forecasts, apparent from both the real-time forecasts and the reforecasts. For wind speed, there is a positive trend in recent years for the real-time forecasts, and no significant trend in the reforecasts.

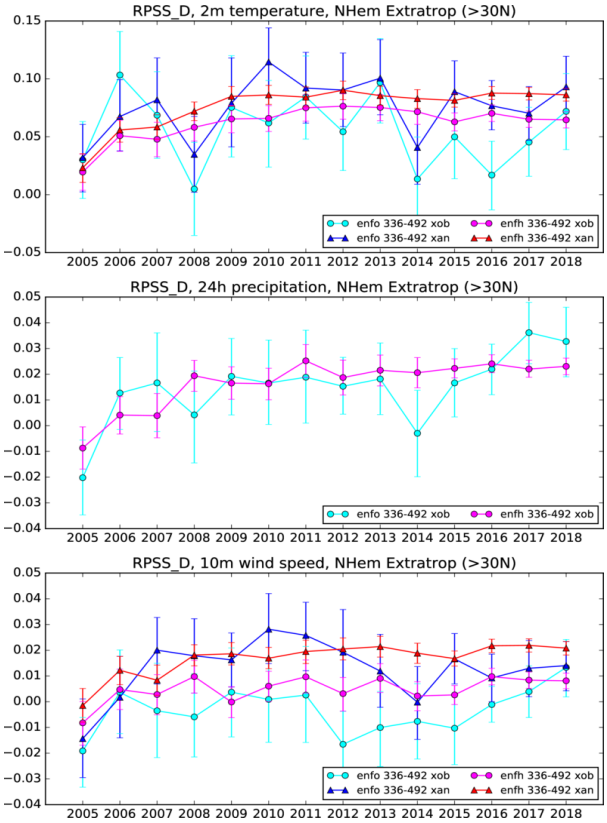

Figure 2: Evolution of the discrete ranked probability skill score of re-forecasts averaged over Week 3 over the Northern Extratropics for the 2-metre temperature (top panel), the precipitation accumulated over 24 h (middle panel) and the 10-metre wind speed (bottom panel). Scores of the reforecasts are drawn as magenta circles (against observations) and red triangles (against ERA Interim analyses), scores of the operational forecasts as light blue circles (against observations) and dark blue triangles (against analyses). The x-axis indicates the year the real-time forecasts were initialized and the year the re-forecasts were produced (but initialized over the previous 20 years).

2.2.2 Scorecard

Figure 3 provides a more extensive comparison of the extended-range forecast skill between the current operational system and the system operational 5 years ago. The fair Continuous Ranked Probability Skill score (CRPSSF) (Ferro et al, 2014) has been applied to remove the dependency on the ensemble size which is present in the CRPSS. It is an estimation of the CRPSS for an infinite ensemble size. The scorecard displayed in Figure 3 shows the difference of CRPSSF between the 2018/2019 and 2013/2014 weekly mean re-forecasts for 20 different variables in the Northern Extratropics and the Tropics. For historical reasons, the verification is based on the weekly periods: days 5-11, 12-18, 19-35 and 26-32 (first calendar week for the forecasts starting on Thursdays 00Z), which will be referred to as week 1, 2, 3 and 4 respectively in this report. According to Figure 3, the skill of the extended-range forecasts has

improved since 2014, particularly in the Tropics where the difference is statistically significant for up to week 4 in the stratosphere and for sea surface temperatures. In the Northern Extratropics, significant improvement is also seen in SST and in the large-scale circulation fields at 200 hPa up to week 4, but in the stratospheric fields the improvement is hardly significance, with a few exceptions. Research re-forecast experiments confirmed that the significant improvement in sea surface temperature, and, as a consequence, in surface temperature, could be attributed to the increase in ocean resolution in November 2016. The experiments also confirmed that a large part of the general improvement in the forecast skill scores, particularly in week 1 in the Tropics, could be attributed to cycle 45R1 and, especially, to changes in the SPPT scheme.

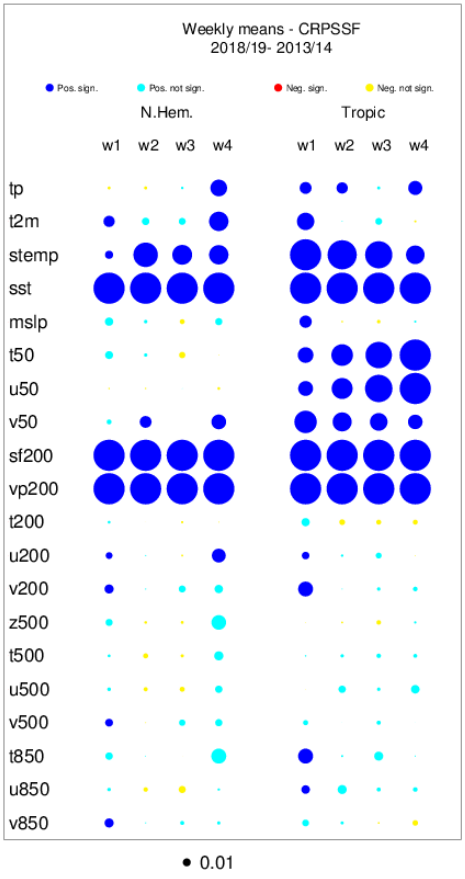

Figure 3: Difference of CRPSSF between the weekly mean re-forecasts produced between November 2018 and March 2019 with the re-forecasts produced between Nov. 2013 and March 2014 and covering the same re-forecast period 1999-2013. The dots represent the difference of CRPSSF for 20 variables and weekly averaged re-forecasts at week 1 (day 5-11), week 2 (day 12-18), week 3 (19-25) and week 4 (day 26-32) over the Northern Extratropics (left side) and the Tropics (right side). The diameter of the dots is proportional to the amplitude of the difference up to a certain value above which the size of the dots remains constant. The cyan/blue (yellow/red) colour indicates that the difference is positive (negative), and therefore the re-forecasts produced in 2018/19 are more (less) skilful than the re-forecasts produced in 2013/14. Blue and red colours indicate that the difference is statistically significant according to a 10,000 bootstrap re-sampling procedure applied to the difference of skill scores. The following variables are verified: total precipitation (tp), 2-m temperature (t2m), soil level 1 temperature (stemp), sea surface temperature (sst), mean sea level pressure (mslp), temperature at 50 hPa (t50), horizontal wind at 50 hPa (u50), meridional wind at 50 hPa (v50), stream function at 200 hPa (sf200), velocity potential at 200 hPa (vp200), temperature at 200 hPa (t200), horizontal wind at 200 hPa (u200), meridional wind at 200 hPa (v200), geopotential height at 500 hPa (z500), temperature at 500 hPa (t500), horizontal wind at 500 hPa (u500),

meridional wind at 500 hPa (v500), temperature at 850 hPa (t850), horizontal wind at 850 hPa (u850), and meridional wind at 850 hPa (v850).

Figure 3 gives a much more optimistic assessment of the progress over the past 5 years than Figure 2 (headline scores). This is due to the fact that the headline scores focus on one single variables (2-metre temperature), one lead time (week 3) and only one region (northern Extratropics). Figure 3 shows that the largest improvements between 2014 and 2019 took place for other variables, regions and lead times. However, Figure 3 still suggests a slight, but not statistically significant, improvement in 2-metre temperature averaged over week 3 over the northern Extratropics, while Figure 2 suggests a slight degradation. This discrepancy can be explained by the fact that Figure 3 provides a fairer comparison by using the same verification period (1999-2012) for the re-forecasts produced in 2013/2014 and 2018/19, while the headline verification is based on a sliding 20-year period (1994-2013 for reforecasts produced in 2014 and 1999-2018 for re-forecasts produced in 2018). As mentioned above, the headline scores are sensitive to the verification period. Ideally, the headline scores should be computed over a fixed period as in Figure 3, but this is not possible because the re-forecast period is sliding by one year every year.

In addition to monitoring the evolution of the probabilistic forecast skill scores in the scorecards, it also important to monitor the predictive skill of sources of sub-seasonal predictability such as the Madden Julian Oscillation (MJO) or sudden stratospheric warmings (SSW). The left panel of Figure 4 indicates that there has been a significant improvement in the daily mean MJO forecast skill scores with a gain of about 3 days of the predictive skill (defined here as the lead time when the bivariate correlation falls to 0.6). Part of this improvement is due to the implementation of a high-resolution ocean (¼ degree) in November 2016. The amplitude of the MJO (not shown) has also improved since 2014. It is still weaker than in ERA Interim, but by 15% instead of 20% in 2014. This strengthening of the MJO was identified in cycle 45R1. The prediction skill of the stratospheric intra-seasonal variability has been estimated by computing the linear correlation between the intra-seasonal variability of an SSW index (zonal wind at 10 hPa and 60°N) during the period November to March 1999-2012. According to the right panel of Figure 4, the forecast skill in the stratosphere is lower in 2018/19 than 5 years earlier, although the difference is not statistically significant. This is not entirely surprising since the 2014 SAC paper showed that the only source of forecast improvement in the stratosphere forecast skill scores between 2004 and 2014 had been increases in the vertical resolution of the stratosphere, in conjunction with a higher top. There has been no change in vertical resolution since 2014.

The prediction of the NAO is of a particular importance for the prediction of European weather. A NAO index has been constructed by projecting the daily 500 hPa height anomalies over the Northern Hemisphere onto a pre-defined NAO pattern. The NAO pattern was defined as the first leading EOF applied to monthly mean 500 hPa height anomalies during the 1950–2000 period. NAO skill scores have been produced by applying the NAO index to the reforecasts and to ERA-Interim and computing the linear correlation between the ensemble means of the reforecasts and ERA-Interim over the extended winter cases (from October to March). The NAO forecast skill scores are slightly higher in the 2018/19 extended winter re-forecasts than in the re-forecasts produced in 2013/14, but the difference is not statistically significant (Not Shown).

All the results presented in this section are based on re-forecast integrations using the same initial conditions (ERA Interim). The impact of improved initial conditions due to better data assimilation system, model and observing system are not taken into account. Therefore, it is likely that the improvements in Figure 2, Figure 3 and Figure 4 are an underestimation of the improvement in the extended-range real-time forecasts.

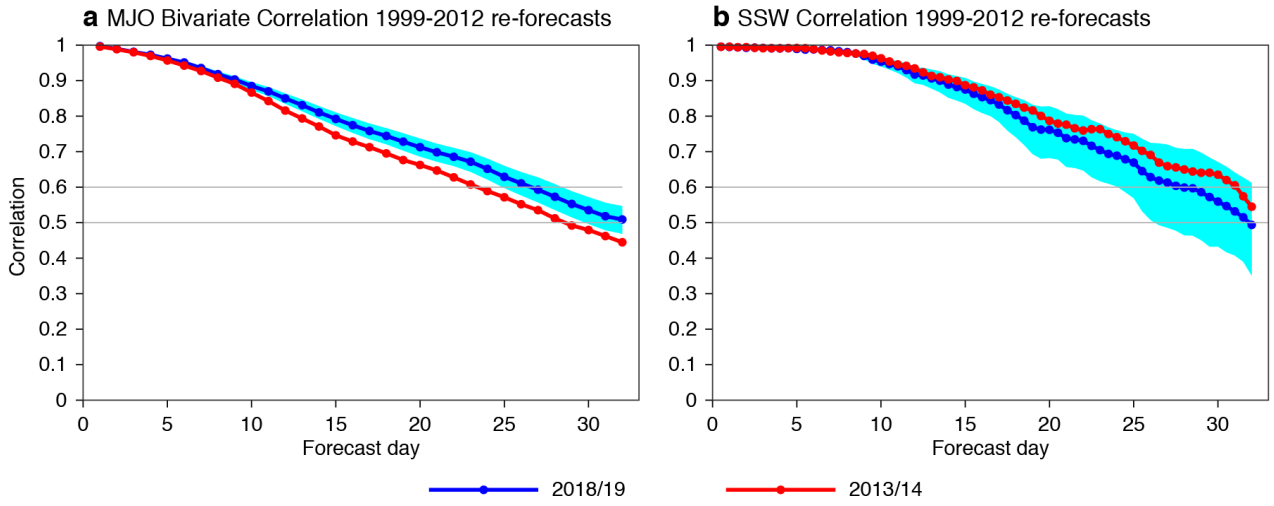

Figure 4: Bivariate correlation between the time series of daily mean MJO RMM1 and RMM2 indices (Wheeler and Hendon, 2004) (left panel) and linear correlation between the time series of the SSW index (right panel) computed from the reforecasts produced from November 2013 and March 2014 (red curve) and from November 2018 and March 2019 (blue curve) as a function of forecast lead time (x-axis). The cyan shaded area corresponds to the 95% level of confidence interval of the blue curve using a 10,000 bootstrap re-sampling procedure.

2.3 Use of ERA5 to initialize the re-forecasts

An important change was introduced with CY46R1: the re-forecasts are initialized from ERA5 instead of ERA Interim. To test the impact of this change, a series of re-forecast experiments have been run initialized from ERA-Interim or from ERA5. The results of these experiments have been documented in an ECMWF Technical Memorandum[3] (TM841, Vitart et al, 2019). A main result is that the skill scores are significantly improved when using ERA5 as initial conditions up to week 3 in the Extratropics and week 4 in the Tropics, except for zonal wind and temperature at 50 hPa in the Tropics, which is slightly degraded, although the difference is not statistically significant (Fig. 5). This can be interpreted as the impact of 10 years of data assimilation system and observing system improvements on extended-range forecasts. The result is not trivial, showing that the impact of atmospheric initial conditions extends to week 3 and week 4 in the Tropics. It highlights the importance of atmospheric initial conditions for obtaining good quality extended-range forecast.

This study also showed that the model surface climatology is more consistent with the real-time forecasts than when using ERA-Interim, which helps removing some known issues in the current operational system (for example wrong cold signal over the central United States in summer). The impact on the real-time forecast skill scores is modest since re-forecasts are used only for calibration. The impact of ERA5 initialization on the EFI skill score is neutral to positive. Since ERA5 is closer to the operational model than ERA-Interim, the scores using ERA5 are likely to be a better estimation of the actual skill of the real-time forecasts.

[3] https://www.ecmwf.int/sites/default/files/elibrary/2019/18872-use-era5-initialize-ensemble-re-forecasts.pdf

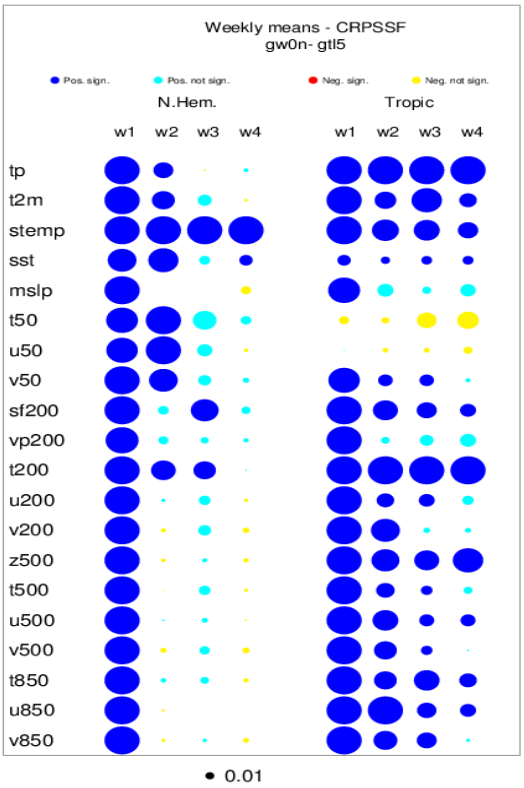

Figure 5: Scorecard of the difference of continuous ranked probabilistic skill scores (CRPSS) between experiment initialized with ERA5 (gw0n) and control (gtl5) over the Northern Extratropics (left column) and the Tropics (right column) for weeks 1 to 4. The size of the dots is proportional to the amplitude of the difference of skill score. The blue (red) colour indicates higher (lower) CRPSS when initializing from ERA5 than from ERA-Interim. Dark blue and dark red colours indicate that the difference is statistically significant at the 1% level of confidence, using a 10,000 resampling bootstrap procedure. The forecasts have been verified against their own re-analysis (ERA5 for gw0n and ERA-Interim for gtl5).

The use of ERA5 for initialization of reforecast comes with another two additional aspects:

- No land surface model simulation needed to initialize the reforecast land conditions
- Use of ERA5 ensemble data assimilation (ERA5 EDA), instead of the operational EDA from recent years. An important advantage of this change is that the ERA5 EDA provides flow-dependent EDA initial perturbations across the re-forecast years instead of the non-flow-dependent perturbations in the current operational set-up.

To put into perspective the impact of 10 years of improvement in the atmospheric initial conditions (ERA5 vs ERA-Interim initialization), the same scorecard as in Figure 5 was produced between the re-forecasts produced operationally in 2018 and 2008, which use the same initial conditions (ERA-Interim) but model versions 10-years apart. Figure 6 shows the skill score (CRPSSF) averaged over the 20 variables of the scorecards for weeks 1 to 4 over the Northern Extratropics (left panel) and the Tropics (right panel) when only the atmospheric initial conditions have changed (orange bars) and when only the model version used in the re-forecasts has changed (green bars). According to Figure 6, the quality of the initial conditions, which also benefit from a better IFS model, has been a main driver for week 1 skill score improvement over the Northern Extratropics. However, in week 4 the improvements are mostly due to the model version used in the re-forecasts. For weeks 2 and 3, the quality of atmospheric

initialization and changes to the model physics have a similar impact over the Northern Extratropics. In the Tropics, the improvements in IFS over the past 10 years have a larger impact than the quality of the atmospheric initialization.

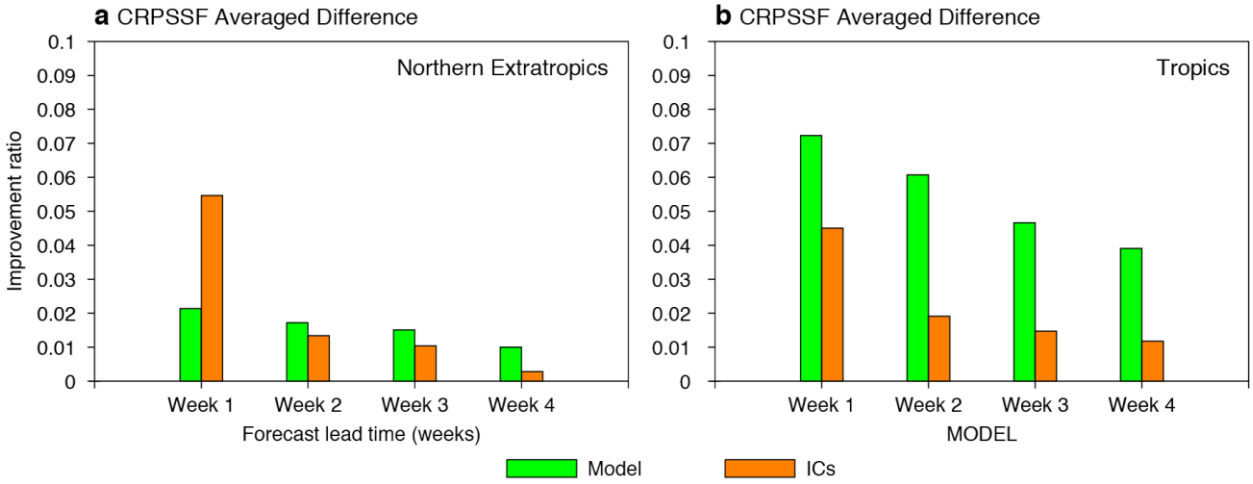

Figure 6: CRPSSF improvement averaged over the 20 variables of the scorecard for weeks 1 to 4 over the Northern Extratropics (left panel) and Tropics (right panel). The green bars represent the difference of CRPSSF between the weekly mean re-forecasts produced in 2018 and 2008 using the same initial conditions (ERA Interim). The orange bars show the improvement when using ERA5 to initialize the re-forecasts instead of ERA-Interim.

2.4 New diagnostics and scores

2.4.1 Prediction of severe weather in Europe

In line with its long-term strategy, ECMWF has recently developed a range of new diagnostics designed to support the prediction of severe weather in Europe. While at medium range, predictions for severe temperature conditions can be directly based on temperature forecast probabilities, at the extended-range, the predictable signal for severe and persistent cold spells is better exploited using large-scale circulation patterns. Ferranti et al (2018) showed that reliable extended-range forecasts of flow patterns such as the NAO and blocking (BL) are instrumental for early warnings of severe cold events over Europe. They used a two-dimensional phase space based on the leading empirical orthogonal functions (EOFs) of mid-tropospheric flow computed over the Euro-Atlantic region (NAO and Blockings). The phase space is an effective tool for monitoring predictions of regime transitions at medium and extended-range, although not all four Euro-Atlantic weather regimes are well represented in this 2-dimensional phase space. Since November 2018, the phase space diagram (Figure 7) has been part of the test products range and is available to the users.

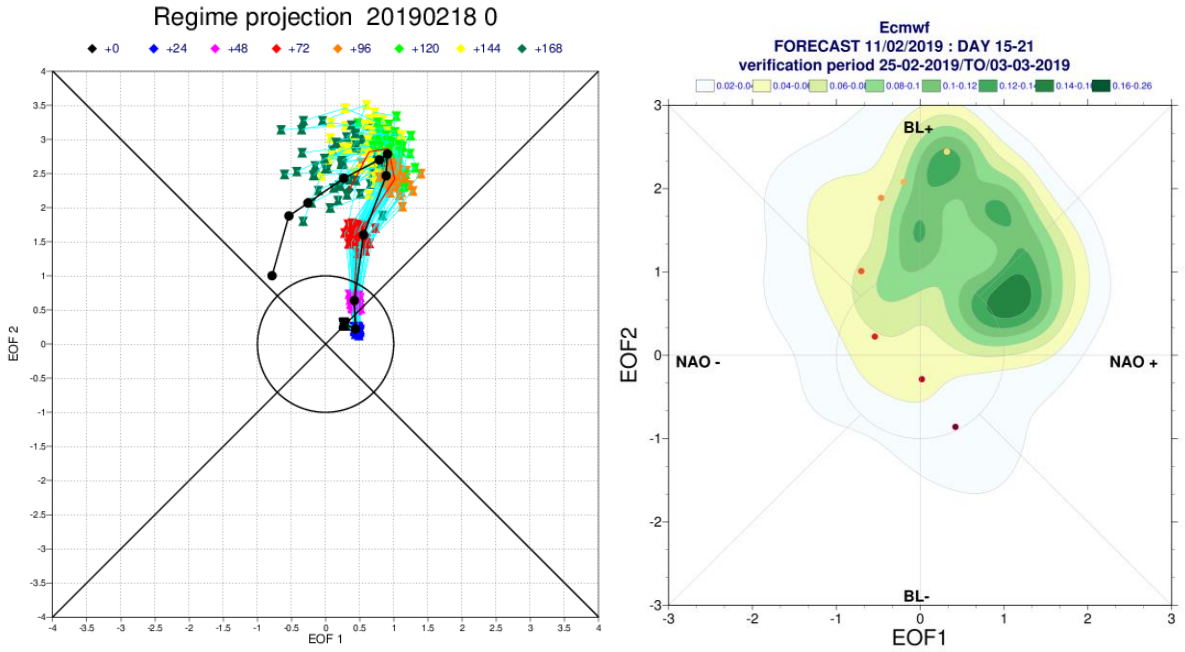

Figure 7: Evolution in regime phase space of the ensemble forecast initialized on 18/02/2019; dark blue (first day) to dark green (last day of the target period). b) Probability density function for forecast initialized on the 11/02/2019. Daily values of the verifying analysis are represented by dots.

Analysis, based on reforecast data, shows that, for these cases, the forecast reproduces faithfully the observed preferential transition paths even at the extended-range. Looking at the MJO impact on the forecast skill, it turns out that forecasts initiated with an MJO event have higher skill and higher reliability in predicting NAO− (see Ferranti et al, 2018). In contrast, the skill of NAO+ prediction is not significantly affected by the existence of an MJO event in the initial state. This asymmetry in the impact of the MJO on forecast skill is consistent with the asymmetry in the strength of the MJO teleconnections discussed by Lin and Brunet (2018) and Yadav and Straus (2017). The skill of the Scandinavian blocking predictions is not highly sensitive to the existence of an MJO in the initial conditions; consistent with this, the blocking exhibits lower predictability than NAO−.

The current skill of the S2S models indicates that sub-seasonal predictions have real potential in providing early warnings of severe cold events over Europe. Current models can deliver skillful predictions for some large-scale patterns 2 weeks ahead, and longer in certain cases (see, for example, Figure 14 in Section 3.3). However, the success of forecasting, weeks ahead, changes in large-scale flow that lead to cold conditions depends on the type of transitions. The ECMWF ensemble, beyond the medium range, can provide reliable probabilities of cold conditions associated with the establishment of Greenland blocking (NAO-). In addition, the predictive skill of such events can be significantly enhanced by MJO activity via tropical-extratropical teleconnections. On the other hand, forecasting probabilities at the extended-range for the occurrence of cold events associated with blocking transitions might present a bigger challenge. Understanding these flow-dependent variations in forecast skill and using the phase space trajectories will enable users to exploit periods of enhanced extended-range predictability.

To complement the medium range clustering products, probabilities of the occurrence of weather regimes have been developed (see example in Figure 8). The distribution of the four Euro-Atlantic flow regimes (NAO+, NAO-, BLO, ATL), as forecasted by the ensemble members at various lead times, is displayed in a cumulative histogram form. The regime attribution is computed on Empirical Orthogonal

Function (EOF) space using "calibrated" anomalies. These anomalies are computed with respect to the re-forecast 20 years climate. A given regime is assigned only if: i) the minimum distance between the anomalies and any of the four regimes is within an "average value" ii) the distance is significantly different from the other regime types. The probabilistic skill for the four regimes is generally higher than the climatological forecast up to 30 days (Ferranti et al, 2018). By day 13-14, the Brier skill scores generally drops below 0.3. For the blocking, the drop occurs a few days earlier. The difference between scores computed with de-biased versus non-de-biased forecasts is small indicating that the mean bias associated with weather regimes is negligible.

In order to monitor the forecast of the strength of the polar vortex and the likelihood of the occurrence of Stratospheric Sudden Warming events (SSW), a new forecast plume product has been designed. SSW events are characterized by a rapid deceleration in stratospheric circumpolar westerly winds and a significant warming of the polar cap region. Following an SSW, there can be an equatorward shift of the tropospheric jet with associated cold conditions in the Northern Hemisphere winter over Northern Europe and warm conditions over north-eastern Canada and Greenland (Baldwin and Dunkerton, 1999). Because extended-range skill is higher when there is an SSW in the forecast initial conditions (Tripathi et al, 2015), SSW events are a potential source of predictability. Figure 9 shows an example of SSW forecast, based on zonal mean zonal wind at 10hPa. According to Figure 9, all the ensemble members initialized on 24 December 2018 predicted the occurrence of an SSW event in early January 2019.

2.4.2 Verification of 46-day extension

An important change to the operational suite has been the extension of the re-forecasts from 32 to 46 days in May 2015 with cycle 41R1. Users can visualize forecasts charts for the extended period from ECCHARTS. At this time range, the model displays low skill in predicting the week to week predictability, therefore products and verification are produced over a 2-weekly period instead. Figure 10 shows the Receiver Operating Characteristic (ROC) Areas for day 33 to 46. According to this figure, the ROC score of 2-metre temperature in the upper tercile is higher over the Northern and Southern Hemispheres in winter (right panel), with ROC areas larger than 0.6 over large portions of Asia and West America. At this time-range, the skill over western Europe is very modest, even in winter. The model displays also significant skill in the tropical regions, like in seasonal forecasting. For precipitation, the skill is limited in the tropical region in all seasons.

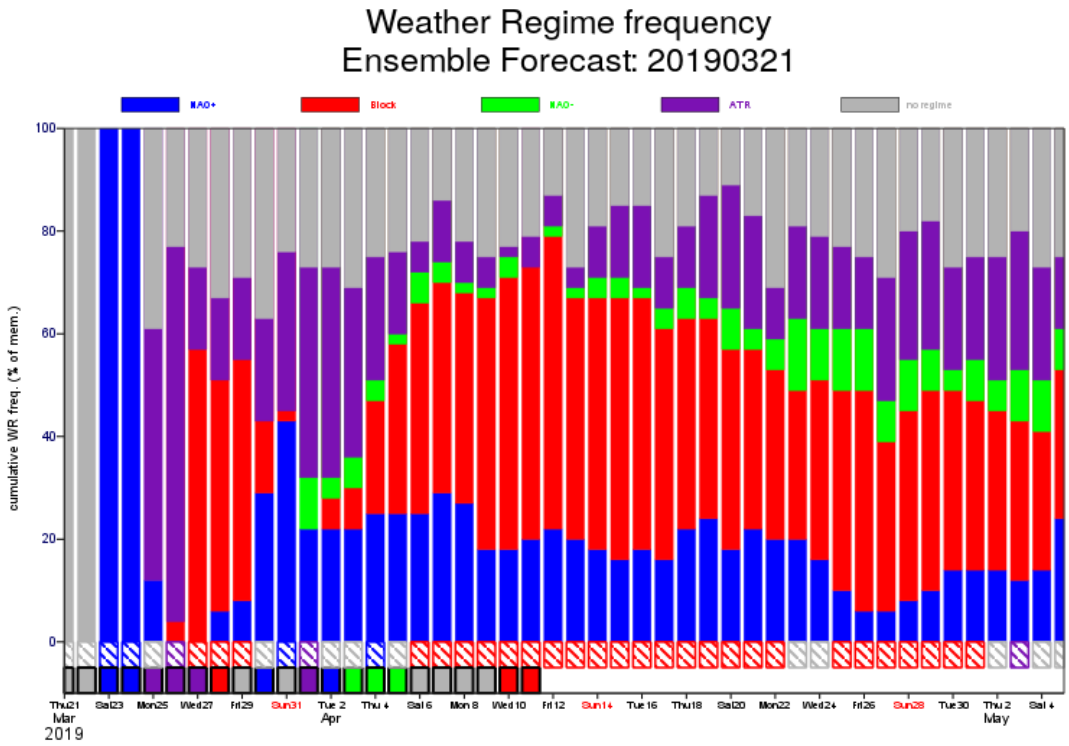

Figure 8: The chart shows the percentage of ensemble forecast members that predict different types of weather regimes for the European domain. In this case, the forecast starting on 21 March 2019 predicted a growing probability of a 'blocking' weather regime from early April, which in fact came to pass on 10 April, according to the verification shown in the small columns at the bottom of the chart. The hatched bars represent the ensemble mean regime attribution.

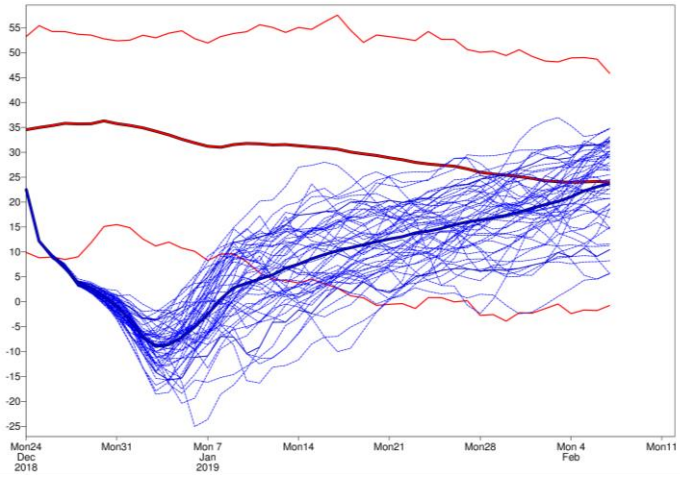

Figure 9: Evolution of the 10hPa zonal mean zonal wind at 60N from the forecast initialized on 24/12/2018. Thin blue lines represent the individual ensemble members, thick blue line the ensemble mean. The red lines indicate the model climatological range: ensemble mean and the 10th and 90th percentiles.

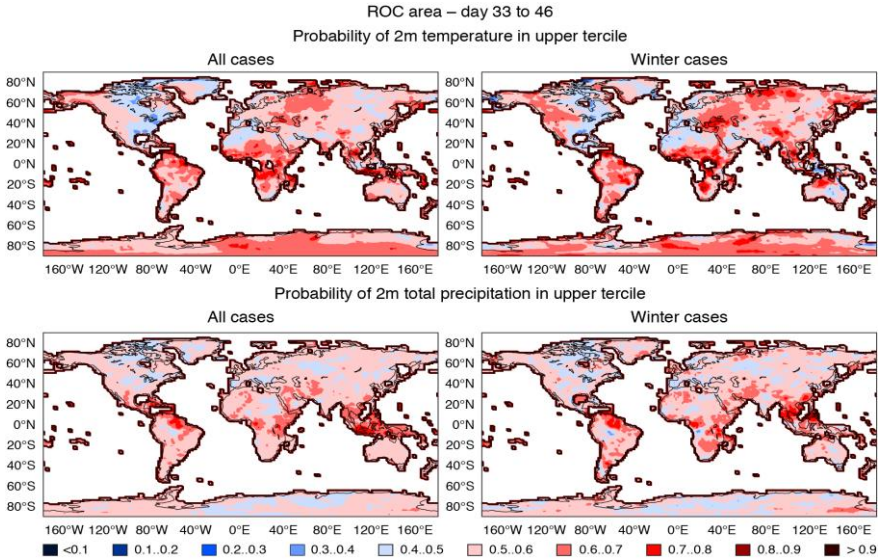

Figure 10: Receiver Operating Characteristic (ROC) Area of the probability of 2-metre temperature in the upper tercile (top panels) and probability of precipitation in the upper tercile (bottom panels) for the lead time averaged between days 33 and 46. The left panels show all the real-time cases between 18 May 2015 and 31 December 2018 (373 cases). The right panels show the extended winter (November to March) cases (152 cases). The blue (red) colour indicates areas where the model has less (more) skill than climatology.

2.4.3 Ensemble spread/skill relationship

An important aspect of ensemble prediction is the ensemble spread which should give a correct indication of model and initial condition uncertainties to produce reliable forecasts. If it is too small, then the ensemble forecast is overconfident; if it is too large, then it is underconfident. A way to measure if the ensemble spread has the correct amplitude is to compare its evolution with forecast lead time to the amplitude of the root mean square error (referred here as spread/skill ratio). Ideally, the two curves should overlap. The spread/skill ratio in the ECMWF extended-range forecasts has been evaluated for the MJO (Vitart, 2017; Lock et al, 2018), but not systematically for the other parameters. An assumption has been that the current operational ensemble generation, which targets medium-range forecasting, should be suitable for weeks 2 to 6 as well. In order to evaluate the spread in extended-range forecasts, a new diagnostic has been developed as (SPREAD/RMSE - 1). This ratio is calculated from the weekly mean anomalies of 20 different variables (the same variables as in the scorecards) and over different regions, so that negative (positive) values correspond to a too small (large) ensemble spread. Figure 11 shows an example from the operational extended-range real-time forecasts produced in 2018.

Figure 11 shows that the ensemble spread of the extended-range forecasts is often in very good agreement with the RMSE up to week 4, particularly in the northern Extratropics. However, the ensemble forecasts are strongly under-dispersive (ensemble spread too small) for the following parameters:

- Surface parameters, such as sea-surface temperature (SSTs) and 2-metre temperature over the Tropics and Northern Extratropics. This indicates that the ensemble perturbations of surface variables are not sufficient to cover the amplitude of the RMSE. For instance, SST perturbations are produced by using only 5 different ocean analyses (1 control + 4 perturbed).
- Zonal winds and temperature at 50 hPa in the Tropics.

More efforts will be dedicated in the coming years to increase the ensemble spread in those areas. Some of these spread/skill ratio errors could be alleviated by including uncertainty in the verifying analysis. In addition, more diagnostics (e.g. flow dependent spread/skill relationship) will be developed.

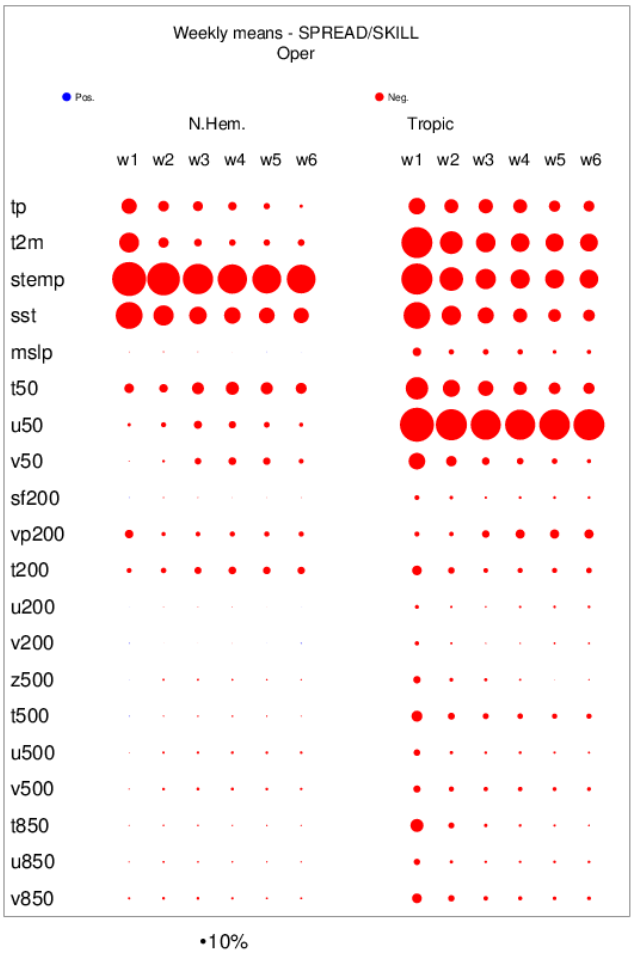

Figure 11: Spread/skill ratio card showing (SPREAD/RMSE - 1) for 20 different variables over the Northern Extratropics (left column) and the Tropics (right column) for weeks 1 to 6. The calculation has been performed on weekly anomalies from all the extended-range real-time forecasts produced in 2018 (105 cases). The diameter of the dots is proportional to the amplitude of the error in spread/skill (the smaller the dot, the better). The red (blue) colour indicates over-confident (under-confident) ensemble forecasts.

3 Comparison with other extended-range forecasts

A key achievement of the World Weather Research Programme (WWRP) and World Research Programme (WCRP) Sub-seasonal to Seasonal Prediction project (S2S) has been the creation of an extensive database containing sub-seasonal (up to 60 days) near real-time (3-week delay) forecasts and reforecasts (Vitart et al, 2017). It follows to some extent the idea behind The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE) database for medium-range forecasts (up to 15 days; Bougeault et al, 2010) and the Climate-System Historical Forecast Project (CHFP) for seasonal forecasts. The S2S database includes near-real-time ensemble forecasts and reforecasts up to 60 days from 11 centres: the Australian Bureau of Meteorology (BoM),

the China Meteorological Administration (CMA), the European Centre for Medium-Range Weather Forecasts (ECMWF), Environment and Climate Change Canada (ECCC), the Institute of Atmospheric Sciences and Climate of the National Research Council (CNR-ISAC), the Hydrometeorological Centre of Russia (HMCR), the Japan Meteorological Agency (JMA), the Korea Meteorological Administration (KMA), Météo-France/Centre National de Recherche Météorologiques (CNRM), the National Centers for Environmental Prediction (NCEP), and the Met Office (UKMO).

The S2S database opened in May 2015. This database represents an important resource for assessing the predictive skill of state-of-the-art extended-range forecast models, investigating predictability and process-oriented diagnostics, but also for comparing the skill of the ECMWF extended-range forecasts with other operational centres. So far more than 80 articles have been published in the peer-reviewed literature analysing and inter-comparing the performance of the S2S Database models for a wide range of topics: MJO, stratosphere, sea-ice, monsoons prediction, extreme events (tropical cyclones, atmospheric rivers, drought, flooding, heat waves, tornadoes...), teleconnections, weather regimes, tropical and extratropical forecast skill. These publications suggest that the ECMWF extended-range forecasts are generally more skilful than the other S2S models for sea-ice cover prediction (Zampieri et al, 2018), MJO prediction (Vitart, 2017), Boreal intra-seasonal oscillation (Jie et al, 2018), prediction of precipitation over various regions (Vigaud et al, 2017), Euro-Atlantic weather regimes (Ferranti et al, 2018), major sudden stratospheric warmings (Taguchi, 2018), tropical cyclone activity (Lee et al, 2018). Several examples, including one where the ECMWF model is not the most skilful, are presented below:

3.1 500 hPa geopotential height skill scores in the Extratropics

Probabilistic forecast skill scores (RPSS, ROC, BSS) have been computed over several regions (Northern Extratropics, Tropics, Southern Extratropics …) for several parameters (500 hPa geopotential, near surface temperature, precipitation ...) using the real-time forecasts in the S2S database between 8 June 2017 and 1 November 2018. During that period, all the 11 S2S models produced real-time forecasts starting the same day of the week (every Thursday). Real-time forecasts were used instead of re-forecasts because of their larger ensemble size and better initialization.

The left panel of Figure 12 suggests that the ECMWF model displays significantly higher forecast skill scores of 500 hPa geopotential height anomalies than the 10 other S2S models. This lead is also visible when using other forecast skill scores (e.g. ROC area), regions and parameters (e.g. precipitation). However, the RPSS is negatively biased for ensemble prediction systems with small ensemble sizes. Weigel et al. (2007) proposed a debiased version of the RPSS which takes into account the ensemble size: the discrete ranked probability skill score (RPSSD). The right panel of Figure 12 shows that ECMWF extended-range forecasts are still significantly more skilful than all the other S2S extended-range forecasts over the Northern Extratropics up to week 3. After week 3, ECMWF and UKMO display similar skill. Over the Tropics, ECMWF maintains a significant lead up to week 4 (not shown). All these results suggest that the well-known lead of ECMWF in medium-range forecasting extends into the sub-seasonal range.

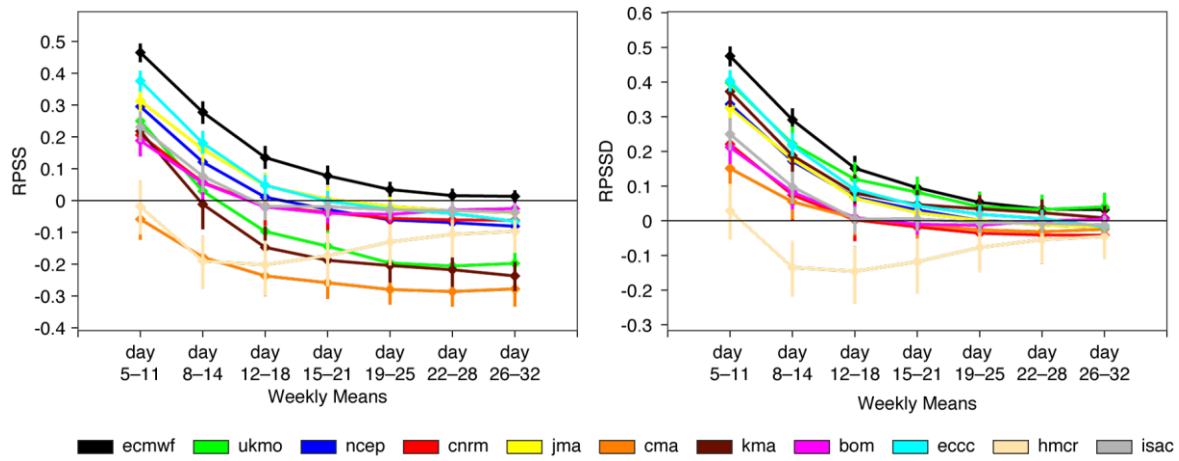

Figure 12: Ranked probability skill scores (RPSS) (left panel) and the RPSSD (right panel) of weekly means of 500 hPa geopotential height anomalies (relative to the common re-forecast period 1999-2010) over the Northern Extratropics. The skill scores have been computed for each model from the real-time forecasts produced once a week between 8 June 2017 and 1 November 2018. The forecast start-dates, used for the skill score computation, are identical for the 11 S2S models (every Thursday).

3.2 The Madden Julian Oscillation

The MJO index computation as described in Gottschalk et al. (2010) has been applied to the reforecasts of ten S2S models (KMA does not provide all the fields needed for the MJO index computation) for the common reforecast period 1999—2010. For each model, the ensemble MJO reforecasts have been verified against ERA-Interim. Figure 13 shows the lead time (in days) when the bivariate correlation (Rashid et al, 2010) reaches 0.6 (defined here as the limit of predictive skill) for the ensemble mean (blue) and control forecast (orange). The statistical significance of the skill scores is measured using a bootstrap resampling technique and the 95% level of confidence is represented by the black bars in Figure 13. According to this figure, the ECMWF MJO extended-range forecasts are significantly more skilful than the MJO forecasts from all the other S2S models. The MJO forecast skill horizon in the ECMWF model is 1 week longer than in the second best S2S model. Part of this lead comes from the larger ensemble size of the ECMWF re-reforecasts (11 members) compared to the other S2S models, except BoM. However, the ECMWF lead subsides when considering only the control forecasts (orange bars in Figure 13), indicating that it is not due to ensemble size only. A similar analysis, using S2S real-time forecasts (as in Section 3.1.) instead of re-forecasts, confirmed this hierarchy between the S2S models (not shown).

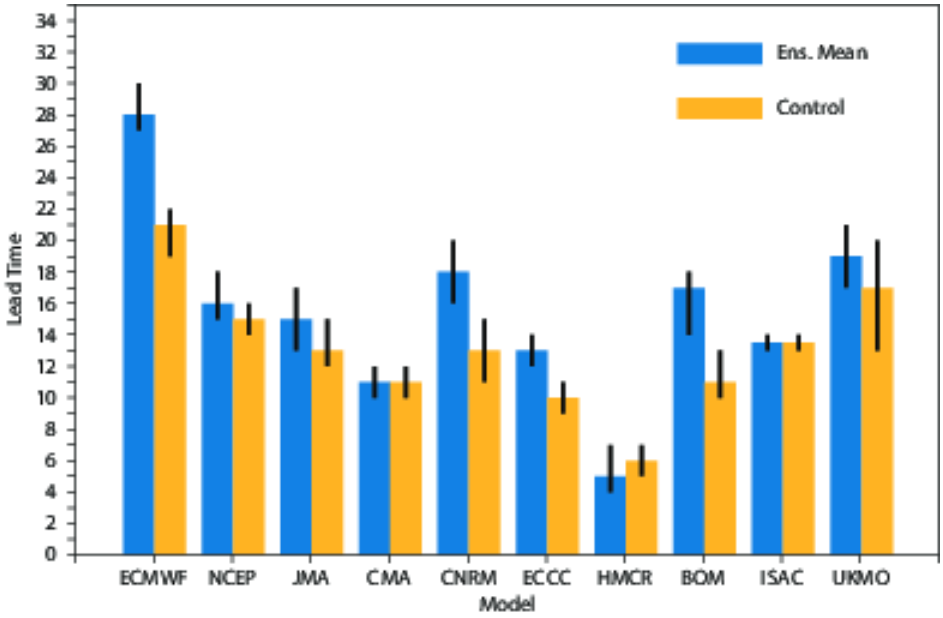

Figure 13: Forecast lead time (in days) when the MJO bivariate correlation between the model ensemble means (blue bars) or control run (orange bars) reaches 0.6. The vertical black bars represent the 95% level of confidence computed from a 10,000 bootstrap re-sampling procedure. From Vitart (2017)

3.3 Euro-Atlantic Weather regimes

Ferranti et al (2018) assessed the skill of several sub-seasonal forecasting systems to predict positive, negative NAO, European blocking (BL) and Atlantic ridge in winter over the period 1999-2010. Depending on the model, the probabilistic skill score drops to zero (a score equal or lower than zero means no better skill than climatology) between day 12 and day 24 for negative NAO, between day 13 and 25 for positive NAO, between day 12 and 18 for Atlantic Ridge and between day 10 and 17 for European blocking (Figure 14). This suggests that extended-range forecasting systems have some skill in predicting these weather regimes more than 10 days in advance, therefore implying potential skill for predicting extreme cold conditions over Europe associated with the negative NAO and BL regime circulations in winter. According to Figure 14, the ECMWF forecasts of the 4 weather regimes are more skilful than the other models during the first 20 days of the extended-range forecast.

3.4 Sea-ice cover

The predictability of sea-ice cover over the Arctic region in the S2S database models which have interactive sea-ice, has been assessed in Zampieri et al (2018). The ECMWF extended-range forecasts display significantly higher skill in predicting the sea-ice edge than the other S2S models (Figure 1 in Zampieri et al. 2018). This is remarkable since the current sea-ice model at ECMWF is LIM2, which is the only sea-ice model among the S2S models that has a simple single-category approach to modelling sea-ice thickness. The results also demonstrate that the current operational system outperforms the previous operational configurations where sea ice cover was persisted and relaxed towards climatology (before cycle 43R1). Another important conclusion from this study was that the ECMWF sea-ice forecasts are skilful up to 46 days, implying that these forecasts could be useful for applications such as ship routing.

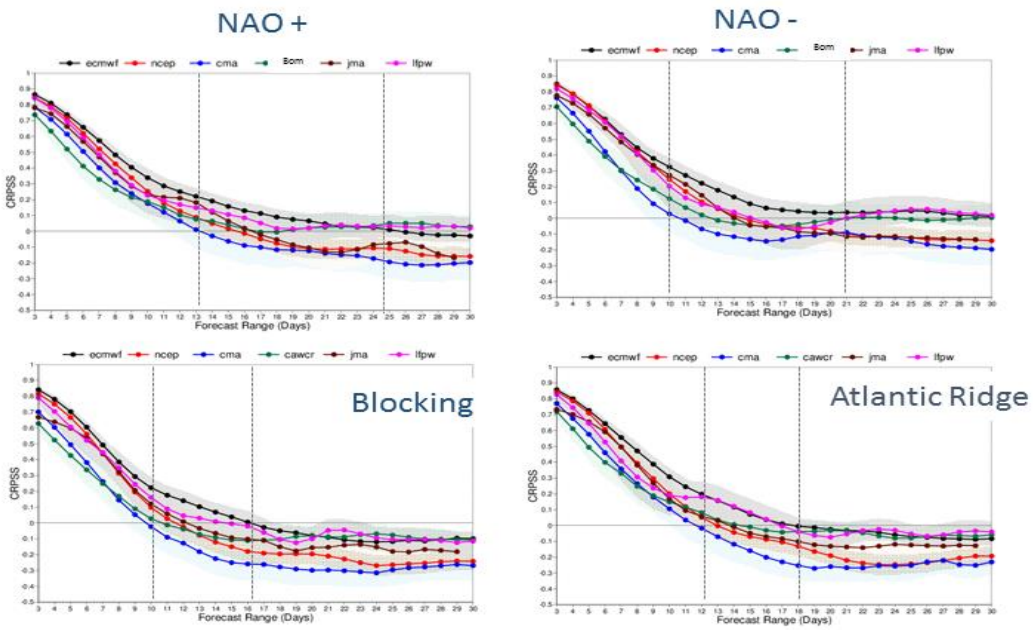

Figure 14: Continuous Ranked probability skill score (CRPSS) of the projection of 500 hPa geopotential height onto the 4 Euro-Atlantic regimes. A 5-days running mean is applied prior the computation of the score. The reference forecast is the climatology (the same period). The scores are computed for the common period of reforecast covering 12 years (1999-2010) from Ferranti et al (2018).

3.5 Sudden stratospheric warmings

The sections above showed examples where the ECMWF extended-range forecasts are more skilful than the other S2S models. However, this is not always the case. For example, the ECMWF extended-range forecasts of a Sudden Stratospheric Warming (SSW) index, (zonal wind anomaly at 10 hPa and 60°N) show a lower correlation with ERA Interim than KMA and UKMO (Figure 15). The difference with KMA is statistically significant. In Figure 15, the S2S models have been ordered by the height of their respective model tops, from the highest top (left side) to the lowest top (right side). Figure 15 suggests a strong relationship between the height of the model top and the SSW index forecast skill, with highest top models having higher skill. The relationship is not as strong if the models are classified by the number of vertical levels (for instance ECMWF and CNRM have more vertical levels than KMA and UKMO, despite a lower top). This result suggests that the prediction of stratospheric intra-seasonal variability at ECMWF would benefit from increased height of the model top or shallower sponge lager, which in the operational forecasts starts already at 30 hPa.

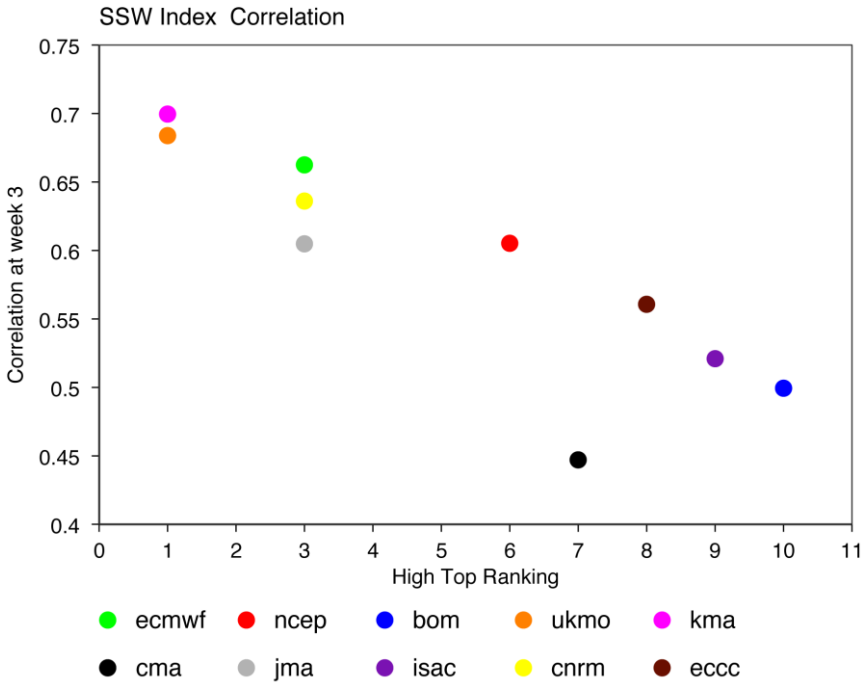

Figure 15: Correlation of the temporal variability of the SSW index averaged over week 3 with ERA -Interim for each S2S model. This verification includes all the real-time forecasts produced between 1 January 2016 and 1 March 2019 during boreal winters (about 120 cases). The models have been classified as a function of the highest vertical level from 0.005 hPa for UKMO and KMA, 0.01 hPa for ECMWF and CNRM to 10 hPa for BOM (x-axis).

4 Limitations of the current extended-range forecasting system

4.1 MJO crossing of the maritime Continent

As reported in the 2014 SAC paper and ECMWF technical memorandum TM738, the ECMWF model often struggles to propagate the MJO across the Maritime Continent. A model inter-comparison using the S2S database suggests that this is an issue common to all the models (Figure 16) and could be linked to the fact that the MJO in the S2S models tends to slow its propagation with lead time (Vitart, 2017). This is an important issue for extended-range forecasts over Europe, since errors in the propagation of the MJO generate errors in the MJO teleconnections and, consequently, on the prediction of the Euro-Atlantic weather regimes.

Kim et al (2016) investigated the characteristics of the MJO propagation across the Maritime Continent in the ECMWF extended-range reforecasts by grouping MJO events initialized in over the Indian Ocean into high- and low-skill events. The authors then analysed the differences in the ocean–atmospheric conditions and differences in the dominant physical processes between the two groups of MJO events. It was found that the relation between initial MJO amplitude and prediction skill is not linear. In particular, the initial distribution of OLR anomalies in high-skill events shows a clear dipole pattern of convection with enhanced convective anomalies over the Indian Ocean and strongly suppressed convective anomalies in the western Pacific Ocean. This dipole mode may support the MJO propagation across the Maritime Continent via the Rossby wave response and associated meridional moist static energy advection. Prominent ocean–atmosphere coupled processes were also correctly simulated during the propagation of high-skill events. However, in low-skill events, the suppressed convective signal over the western Pacific was almost absent and less organized, and the ocean–atmosphere coupled processes

were not simulated correctly. It was found that in both high- and low-skill events, the amplitude of the convective anomaly decreased significantly after about day 15, possibly due to the systematic mean model bias. Based on this analysis, Kim et al (2016) concluded that the strong wet bias in the vicinity of the Maritime Continent makes the west Pacific area unfavourable for MJO propagation. This bias could explain the difficulty of the model to propagate the MJO across the Maritime Continent, thus limiting its prediction skill.

The wet biases over the eastern part of the Maritime Continent are significantly reduced when the IFS is run in uncoupled mode, forced by observed SSTs. A series of ensemble re-forecasts (15 members covering the period 1989-2016 once a month) with the atmospheric model (CY45R1) has been run forced by observed SSTs. MJO diagnostics applied to these experiments indicate that the percentage of strong MJO events propagating from Phase 2 (Indian Ocean) to Phase 6 (west Pacific) is statistically significantly higher in the uncoupled experiments than in the coupled experiments (43% compared to 30%). 50% of observed strong MJOs propagate from Phase 2 to Phase 6-7. The difference becomes significant only during the transition from Phase 5 to Phase 6 and increases with lead time. This result confirms that the SST biases play a role. This suggests that, although the ocean-atmosphere coupling improves significantly the MJO prediction skill overall (e.g. Woolnough et al. 2006), thanks mostly to a more realistic MJO propagation over the Indian Ocean and west Pacific, some aspects of the MJO propagation may be degraded due to systematic errors introduced by the coupling.

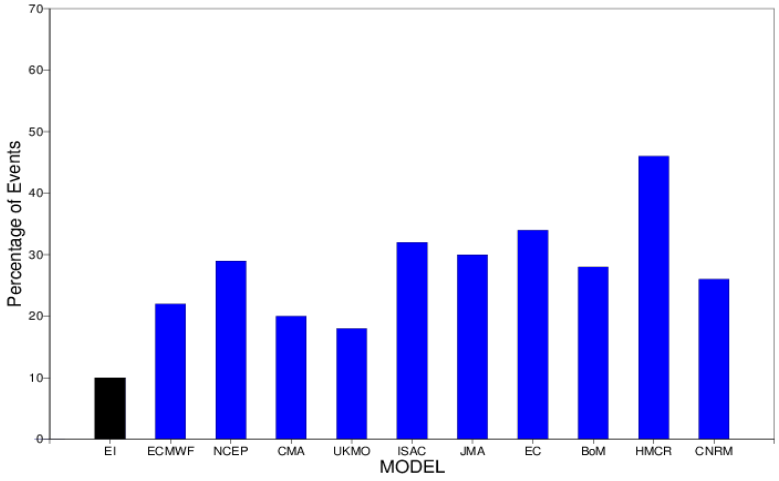

Figure 16: Percentage of MJO events which are located in Phase 2 or 3 (active phase over the Indian Ocean) in the initial condition with an amplitude larger than 1, which never propagate into the western Pacific (Phases 6 or 7) even as a weak MJO during the following 30 days. The black bar corresponds to ERA Interim.

4.2 MJO teleconnections

Another important issue of the ECMWF extended-range forecasts is the underestimation of the MJO teleconnections over the Euro-Atlantic region. Using reanalysis data, Cassou (2008) showed that there was a link in re-analysis data between the MJO and North Atlantic Oscillation (NAO) during boreal winter (DJF). The probability of a positive phase of the NAO is significantly increased about 10 days after the MJO is in Phase 3 (Phase 3 + 10 days), and significantly decreased about 10 days after the MJO

is in Phase 6 (Phase 6 + 10 days). The probability of a negative phase of the NAO is decreased (increased) about 10 days after the MJO is in Phase 3 (Phase 6). However, the ECMWF model underestimates the impact of the MJO on the NAO by more than 50%. Unfortunately, there has not been any significant improvements in the MJO teleconnections since 2014 (Figure 17). Although there has been a slight improvement over the North Pacific (weaker teleconnections), the teleconnections over the Euro-Atlantic sector are slightly weaker in 2018/19 than in 2013/14.

The same diagnostics were applied to all the S2S models in Vitart (2017). The results indicated that all S2S models underestimate the amplitude of the MJO teleconnections over the Euro Atlantic region and that this underestimation is statistically significant. NCEP produces the most realistic MJO teleconnections, followed by ECMWF and UKMO. This study also showed that higher atmospheric resolution S2S models produce generally stronger MJO teleconnections than low resolution models. However, there are exceptions, such as NCEP which produces stronger teleconnections than higher resolution models. This is an important issue for sub-seasonal prediction since the S2S models are currently able to exploit only a small portion of the predictability associated to the MJO. This suggests also that there is scope for large extended-range forecast skill improvement over Europe.

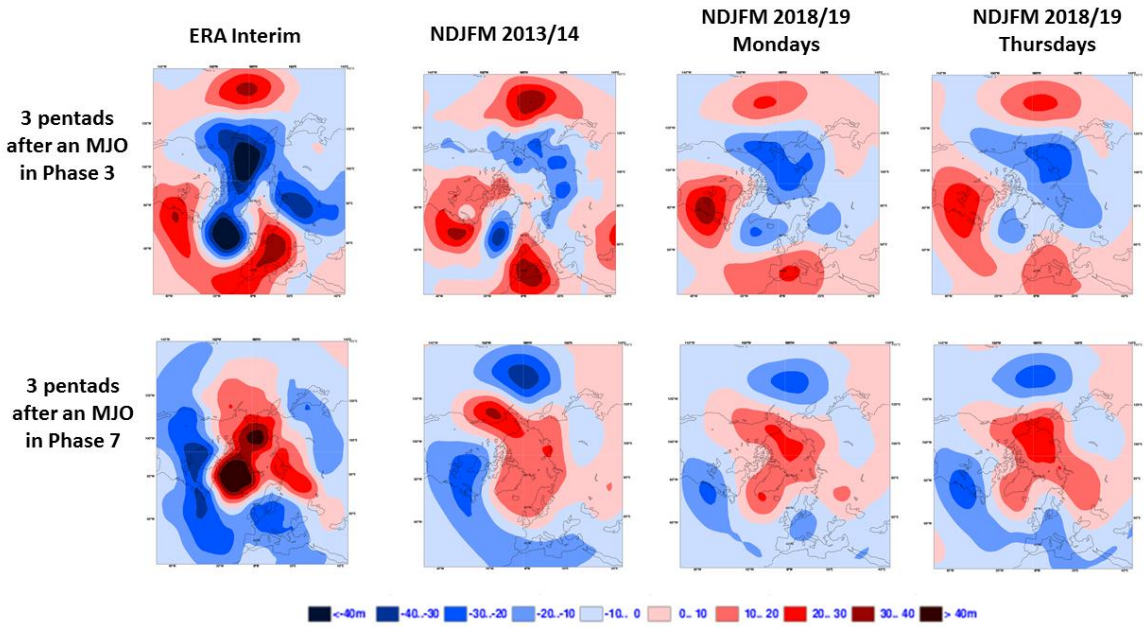

Figure 17: Composites of 500 hPa geopotential height 3 pentads after an MJO in Phase 3 (top panels) or Phase 7 (bottom panels) from ERA interim (left panels), the re-forecasts generated from November 2013 to March 2014 (second column), the re-forecasts generated every Monday from November 2018 to March 2019 (third column), the reforecasts generated every Thursday from November 2018 to March 2019 (right panels). The re-forecasts cover the common period 1999-2012 and only 5 ensemble members have been used for consistency.

This lack in teleconnections could be caused by:

1. Errors in the representation of the MJO
2. Errors in the tropical Rossby wave response and extratropical Rossby wave propagation
3. Errors in the stratospheric response and teleconnections

Vitart (2017) showed that the MJO teleconnections become weaker as lead time increases suggesting that model systematic errors could be responsible. In order to assess the impact of errors in the Tropics, and most especially in the representation of the MJO, a set of 15-member re-forecasts using cycle 45r1, has been produced over the period 1989-2016. In these re-forecasts, the Tropics (20S-20N) have been relaxed towards ERA Interim, producing "perfect" forecasts of the tropical circulation, including the MJO. The right panel in Figure 18 shows that the MJO teleconnections obtained 3 pentads after an MJO in phase 3 are still significantly weaker than in ERA interim (left panel of Figure 18) and not significantly stronger than in the Control experiment (middle panel of Figure 18). This important result suggests that the origin of this problem is not in the Tropics, but most likely originates from systematic errors in the Extratropics.

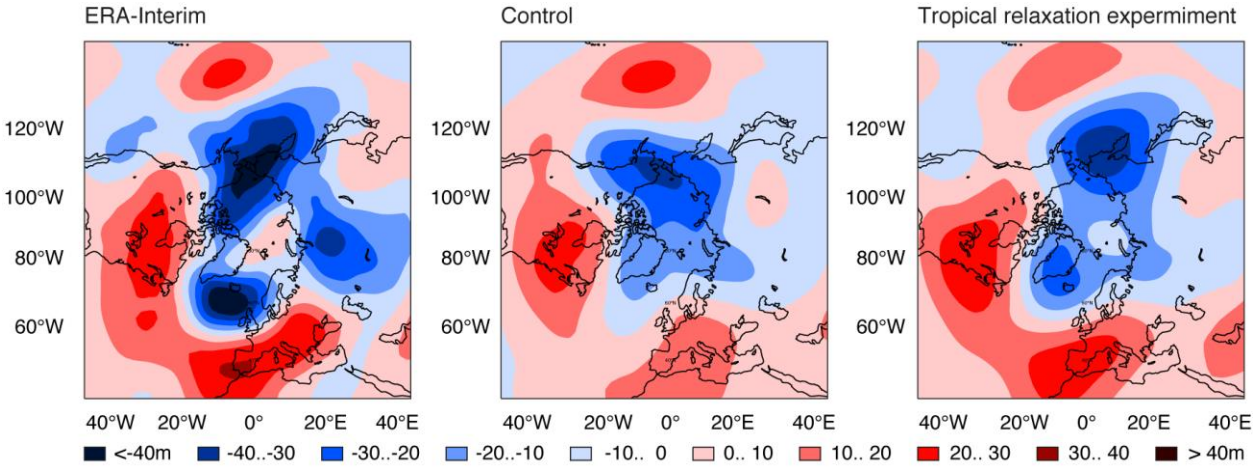

Figure 18: Composites of 500 hPa geopotential height anomalies 11—15 days (third pentad) after a strong MJO (amplitude larger than 1) in Phase 3 (active phase of the MJO over the East Indian Ocean) in ERA Interim (left panel), Control experiment (middle panel) and experiment where the tropical band 20S-20N has been relaxed towards ERA Interim (right panel).

4.3 Rossby wave packets

The impact of the MJO on the Euro-Atlantic weather regimes occurs through the generation and propagation of Rossby waves (e.g. Cassou, 2008). The propagation of Rossby waves in the Extratropics has been diagnosed recently in the ECMWF model and in the other S2S models by Quinting and Vitart (2019) by applying the Rossby wave packet diagnostic (Grazzini and Vitart, 2015) to the model's reforecasts over the common period 1999-2010. Quinting and Vitart (2017) showed that all S2S models, except 2, slightly underestimate the RWP initiation frequency over the central North Pacific (180°W to 140°W) (their Figure 2). This small, but statistically significant, underestimation appears to be rather independent of the horizontal grid spacing as it occurs in the BoM (about 200 km resolution) as well as in the ECMWF (about 18/36 km resolution) model. The biases in the decay of the Rossby wave packets display a strong dipole structure in all the S2S models: negative bias over western Europe and positive bias over Eurasia. This clearly indicates that the RWPs do propagate too far east. Although all S2S models are affected, the amplitude of this dipole is reduced with increased horizontal resolution. The dipole in the RWP decay frequency bias over the Atlantic-European sector indicates that the models

may have difficulties in representing the formation of atmospheric blocks in this region which halt the propagation of RWPs.

4.4 Ocean biases

In spite of the benefits of ocean-atmosphere coupling for predictions of phenomena such as the MJO (e.g. Woolnough et al, 2007), ocean-atmosphere coupled forecasts display large systematic errors in SSTs, which can exceed a few degrees after 4 weeks. In regions where the atmosphere is responsive to SST values, these systematic errors in SSTs will affect the atmospheric circulation and may degrade the forecast skill scores. Middle latitude SST biases, especially those related to sharp SST fronts, have also received increasing attention in recent years. Studies based on AMIP experiments (e.g. Nakamura et al, 2008, Woollings et al, 2010, Keeley et al, 2012) have documented the impact of errors in the representation of the westerly currents (Kuroshio, Gulf Stream) on the position of the jet stream and storm tracks. Scaife et al. (2011) found that the improved representation of SSTs over the North Atlantic Current, resulting from the increasing in ocean resolution from 100 to 25 km, led to improved Atlantic winter blocking in the UK Met Office coupled climate model. More recently, Lee et al (2018) ran an AMIP experiment where the SSTs over the Gulf Stream were taken from a coupled experiment with a low-resolution ocean (~100 km grid spacing), which displayed large SST errors. Results indicated a significant impact on the position of the jet stream.

Re-forecast experiments have been performed starting on the 1st and 15th of each month from November to March 1989 to 2015 (270 start dates). The control experiment is the coupled experiment with cycle 43R1 at a Tco399L91 resolution. The ocean component is NEMO with a ¼ degree resolution. Figure 19 illustrates the SST biases obtained at week 4 (day 26-32). Compared to the SST from ERA Interim (a blend of SST analyses from NCEP before 2009 and high-resolution OSTIA after 2009), the coupled model develops large scale biases, which have the potential for influencing the large-scale atmospheric circulation. Figure 19 shows that the SST biases result in increased meridional SST gradients (warmer Tropics and colder Extratropics); strong warm biases appear over stratocumulus areas off the American and African coast, and over the Southern Ocean; there are also large-scale warm biases over the Indian Ocean, with possible consequence for the atmospheric convection. Aside from these large-scale biases, narrower but strong biases (larger than 2^{o}C) are visible over the sharp SST fronts and Western boundary currents such as the Gulf Stream and Kuroshio. The latter are common to most coupled general circulation models (GCMs) and are related to the insufficient resolution of the ocean. It is estimated that resolutions finer than 1/12 of degree are needed to realistically represent the Gulf Stream separation (Hewitt et al, 2016).

In order to assess the impact of these SST biases on the extended-range forecasts, a series of experiments has been set up where the ocean sees the atmospheric fluxes in the same way as in the control experiment, but the SSTs from the ocean model are bias corrected before they are passed to the atmosphere. The correction consists of removing the SST biases estimated from the Control experiment, which depend on the forecast lead-time and calendar starting date. In the first experiment, the bias correction has been applied globally (BC_GL). In a second set of experiments it has been applied only over the Extratropics (North of 30N and South of 30S). This experiment will thereafter be referred to as BC_ET. In a third experiment, the bias correction has been applied only over the North Atlantic (BC_AT). Although crude, this methodology appeared efficient in removing SST biases in the regions where the corrections were applied, except in the Tropics for BC_GL.

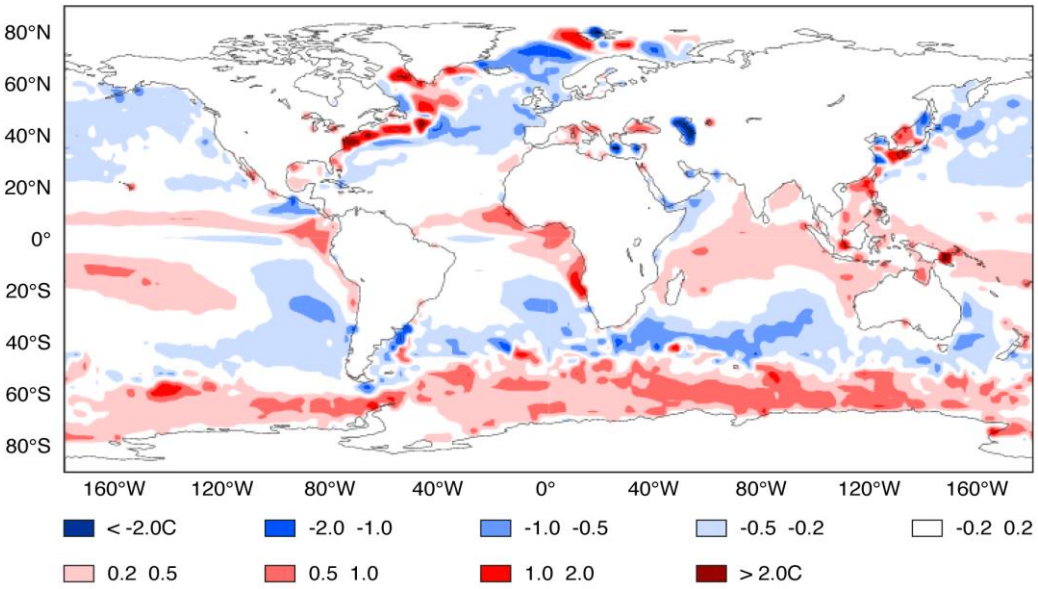

Figure 19: SST biases (relative to ERA-Interim) at the time range of 26-32 days for the period NDJFM 1989-2016.

Figure 20 suggests that the European skill scores are improved when using SST bias correction over the Northern Extratropics (BC_ET) in weeks 2, 3 and 4 when there is a strong MJO in the initial conditions. On the other hand, the SST bias corrections do not impact significantly the European skill scores when there is no MJO in the initial conditions (Figure 20, right panel). BC_ET also displays higher NAO skill scores in the extended-range than the Control experiment, although the difference is statistically significant only after 28 days. The two other SST bias corrected experiments also produced similar improvements (not shown), suggesting that it is the corrections in the northern Atlantic which have an impact.

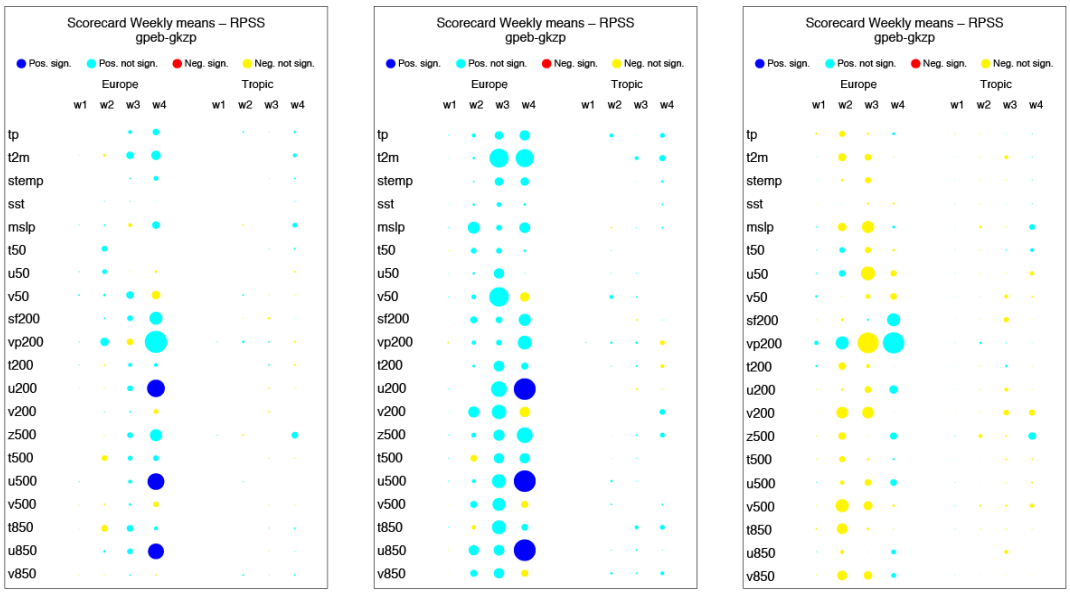

Figure 20: Scorecards of BC_ET CRPSS over Europe and Tropics relative to Control when all the forecasts are included (left panel), only cases with a strong MJO in the initial conditions (middle panel) and when there is no MJO in the initial conditions (right panel).

In order to assess the impact of the SST biases on MJO teleconnections, the amplitude of the MJO teleconnections produced by the 4 experiments has been compared. Results indicate that the amplitude of the MJO teleconnections in the Control experiment is significantly weaker than in ERA Interim with only 50% of the amplitude after an MJO in Phase 3 and 30% after an MJO in Phase 7 (Figure 21). The amplitude of the teleconnections after an MJO in Phase 3 is about the same in the three SST bias correction experiments as in Control (Figure 21, left panel). This suggests that the SST biases do not affect the impact of the MJO on the positive phase of the NAO. However, the amplitude of the MJO Phase 7 teleconnections increases from 30% in control to about 45% (relative to ERA Interim) in the three SST bias correction experiments, which all show remarkably similar statistics (Figure 21, right panel). This difference between the SST bias correction experiments and Control is statistically significant. Since the teleconnections after an MJO in Phase 7 project onto a negative NAO, this result indicates that the SST biases over the North Atlantic impact the probability of negative NAO associated with the Madden Julian Oscillation. This may explain why the uncoupled models of the S2S database (e.g. ECCC and JMA) display stronger and more realistic teleconnections over the Euro-Atlantic sector after an MJO in Phase 7 than all the coupled S2S models (Vitart, 2017), which display similar SST biases over the Northern Atlantic.

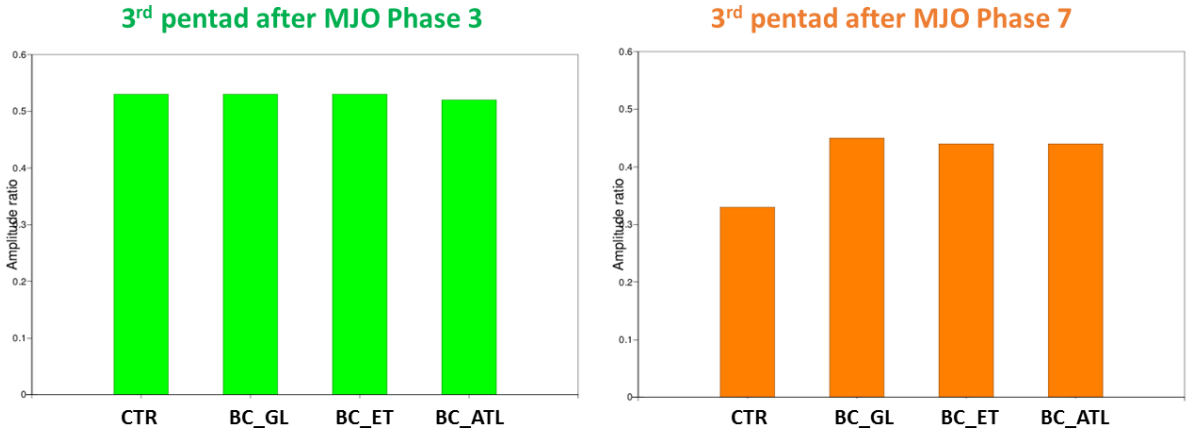

Figure 21: MJO teleconnection amplitude ratio compared to ERA Interim. The absolute values of composites of geopotential height anomalies 11 to 15 days after an MJO in Phase 3 (left panel) and after an MJO in Phase 7 (right panel) have been averaged over all the grid points North of 60N for each experiment (CONTROL, BC_GL, BC_ET and BC_AT).

Overall, these experiments suggest that most of the impact of the SST biases comes from the North Atlantic, most likely from the errors in the position of the Gulf Stream. Further experiments and diagnostic work are needed to fully prove this fact in the ECMWF forecasting system. Although the impact of the Gulf Stream SST biases has not been strictly isolated, the results presented here are consistent with previous findings in seasonal and climate models at other time-scales, which show that correcting the Gulf Stream errors impact the European winter blocking frequency (Scaife et al, 2011, Keeley et al, 2012). More details on these experiments can be found in Vitart and Balmaseda (2018).

The impact of the SST biases on overall forecast skill is not huge, but it is large enough to make the correction of these systematic error an important contribution to improved extended-range forecasts. The best solution would be to reduce the SST biases in the coupled model. However, if the relevant SST errors are related to the position of the Gulf Stream, there are no immediate prospects for reducing them. The errors tend to improve with higher oceanic resolution. Experimentation with a 1/12 degrees version of NEMO will be performed to determine if this resolution helps to reduce these biases. However, even

if this is the case, it is very unlikely that such resolution will become operational in the next 5 years. Another option would be to artificially fix this problem, by correcting the SSTs used by the atmospheric model, similarly to the SST partial coupling fix used for the medium-range. The bias correction investigated in this study operates only at the surface, which, although simple, is not ideal. Besides, it would be cumbersome to use it operationally, since a re-forecast dataset needs to be run beforehand to estimate the systematic errors, followed by a second set of re-forecasts with bias corrected SSTs. There are other reasons against the use of surface flux correction in coupled models, especially in the context of non-initialized integrations: in the long term, it can slow progress in model development, since it masks issues and interfere with coupled feedbacks. On the other hand, from the perspective of initialized forecast, it would allow ECMWF to produce more reliable and skilful extended-range forecasts over Europe. It may be possible to find a consistent framework for treatment of model error by making use of assimilation terms to estimate the bias terms more efficiently, which would provide a continuous transition from assimilation to forecast model.

5 Case study: wintertime forecast bust

In this section, a case study is presented in order to illustrate some of the issues mentioned above. We discuss here the temperature evolution during January and February 2019. The temperature over Europe in early winter was generally mild, until around the middle of January we observed a cold spell that lasted for about 2 weeks followed by a warm period that lasted until the first week of March. Figure 22 shows the ensemble distribution of 2m temperatures averaged over Northern Europe at 15-21 days. At this time range, the forecast performance for the period in question was generally good: the ensemble mean captured the observed variations in 2m temperature, and the ensemble spread also showed variations associated with predictability: i) generally smaller spread than the model climatology, and ii) increased spread during the transition periods, such as the 14 and 28 of January.

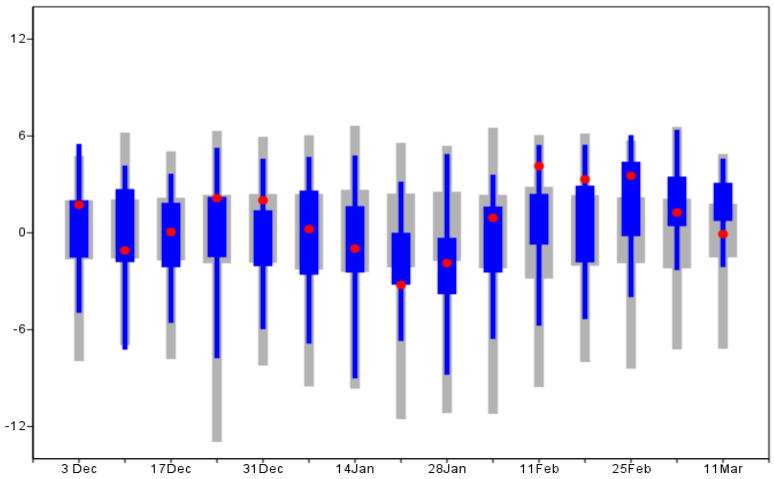

Figure 22: Weekly mean 2m temperature anomalies averaged over Northern Europe (48-70N 10W-30E). The red circles represent the analysis, the 15-21 day forecast and re-forecast distributions are represented by the blue and grey box - whiskers.

While the cold spells during January were successfully predicted in the extended-range, the forecast of warm period starting in mid-February showed contrasting behaviour at different lead times. Figure 23 shows the 2-metre temperature ensemble mean anomalies valid for the week starting on the 11 of

February 2019, when the transition from cold to warm occurred for different forecast ranges (although hereafter the ensemble mean is used to identify the predictable signal, our analysis has included an inspection of the full forecast distribution). The forecast initialized on 28 January 2019 provides the very first indication of a temperature change, while the forecast issued 4 days earlier (24 January 2019) maintained unusually large amplitude cold anomalies, uncharacteristic of the ensemble mean at week 4 forecast range, and produced a large forecast bust in the Northern Hemisphere skill scores for week 4. We are interested in analysing these two predictions to explore the possible reasons for such different behaviour.

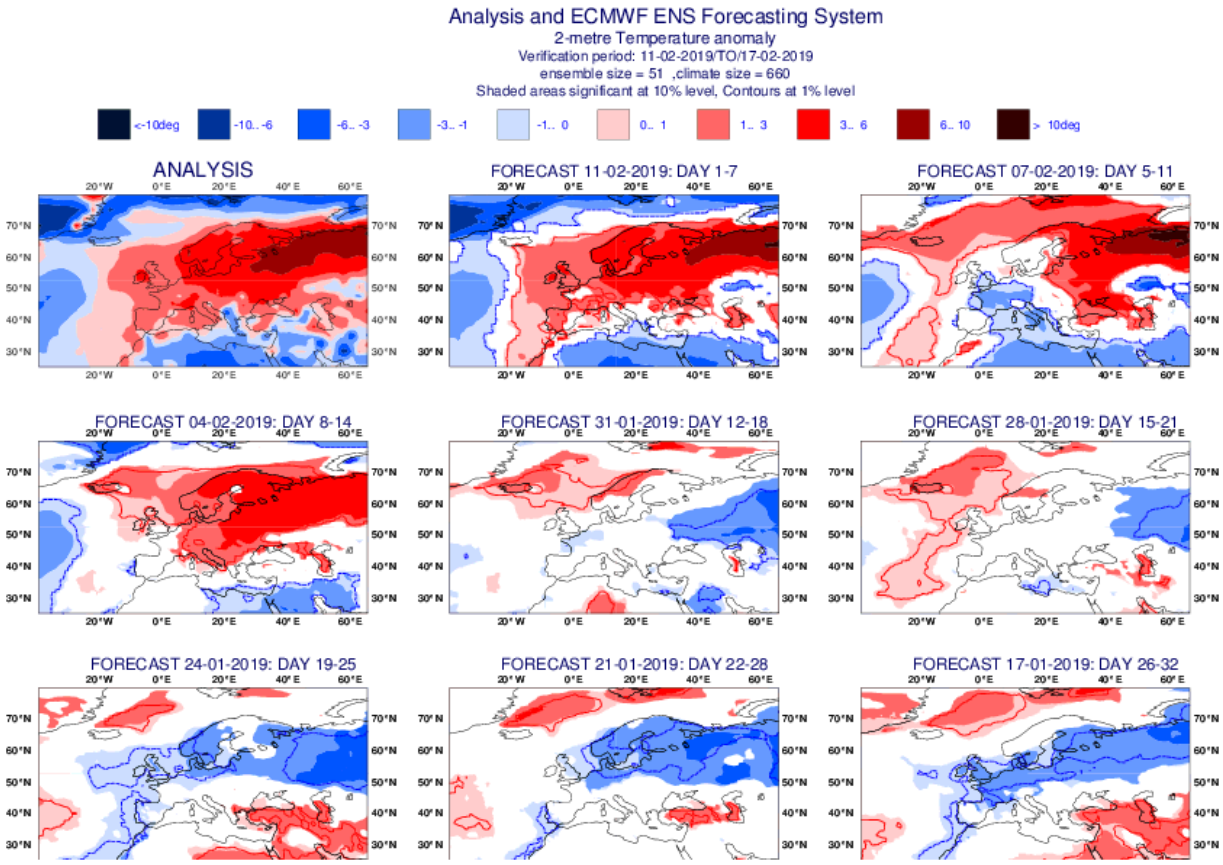

Figure 23: Weekly mean anomalies relative to the past 20-year climate. The first panel corresponds to the anomalies computed using ECMWF operational analysis and reanalysis for a given week. The other panels correspond to the eight forecasts starting half a week apart and verifying on that week. The model anomalies are relative to the model climate computed from the model re-forecasts. The areas where the ensemble forecast is not significantly different from the ensemble climatology according to Wilcoxon-Mann-Whitney (WMW)-test are blanked. The time range of the forecasts is day 1-7, day 5-11, day 8-14, day 12-18, day 15-21, day 19-25, day 22-28 and day 26-32. This figure gives an idea of how well the predicted anomalies verified against the ECMWF analysis and about the consistency between the monthly forecasts from one week to another.

Consistent with the temperature predictions, the circulation associated with the onset of warm anomalies over Europe was predicted by the 28 January 2019 forecast but was incorrectly represented by the 24 January 2019 forecast (Figure 24). The 28 January 2019 forecast, despite being initialized in the presence of a negative phase of the NAO (NAO-), was able to correctly represent the anomalous high over Europe and captured the negative phase of the Pacific-North America (PNA) pattern. In contrast

the 24 January 2019 and 21 January 2019 forecasts exhibited a flow configuration like the NAO-, with low/high anomalies over Southern Europe/Iceland, and a dominant PNA+, in phase opposition to the one in the verification. Those forecasts exhibited a wave number one structure over the Northern Hemisphere reminiscent of the Arctic Oscillation pattern. Figure 24 suggests that the temperature change over Europe is associated with a transition between hemispheric circulation regimes. Inspection of global temperature predictions from S2S forecast confirms that neither the cold anomaly over West coast of North America nor the warm anomaly over Europe were predicted beyond 3 weeks. In addition, we noticed a strong consistency among the S2S forecast errors at 3.5-week range suggesting a possible existence of a common driver.

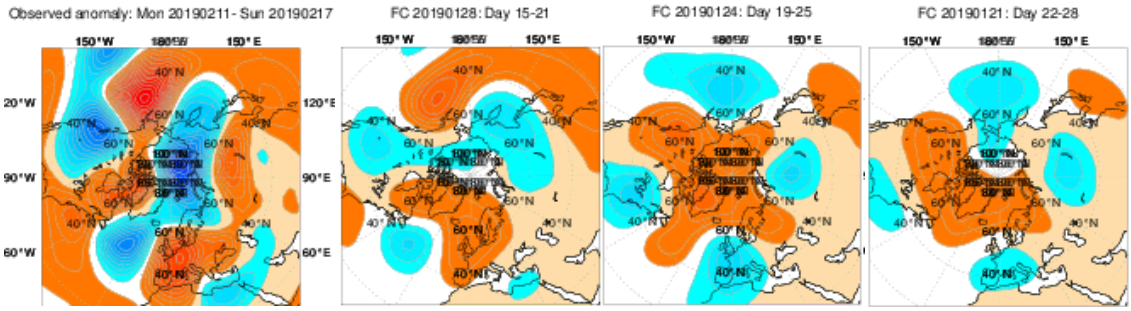

Figure 24: Geopotential height weekly mean anomalies at 500hPa: a) analysis b) c) d) ensemble mean from 28/01/2019, 24/01/2019 and 21/01/2019 forecasts. The analysis shows a PNA- and NAO+. Forecast beyond week 3 (panels c and d) show the opposite phases.

The tropical forcing can favour the occurrence of certain phases of the PNA and NAO. Thus, PNA- (+) lags the MJO active convection over the Indian (Western Pacific) by about one week. Around the middle of January, tropical organized convection over the Indian Ocean, associated with an MJO event, was propagating eastward. By the end of January, the enhanced convection reached the West Tropical Pacific favouring the occurrence of the PNA+ and NAO-. In February, tropical convection, rather than propagating further east, remained stationary over the date line for some time. This stationary enhanced convection over the date line was well represented by all forecasts, even the ones beyond week 3, which can explain the associated PNA+/NAO- response in the forecast beyond week 3, and the unusual strong intensity of the forecast anomalies that resulted in a forecast bust.

The forecast beyond week 3 exhibit the canonical MJO teleconnection over the dateline with PNA+ and NAO-, but the observations (and the week 3 forecast) deviate from that canonical response, showing instead a PNA- (Figure 24). The mechanisms behind the observed PNA- are not understood and it is not clear if they were predictable 3 weeks in advance. They may be related to Rossby wave sources in mid latitudes (Seo and Lee, 2017), interference of tropically forced Rossby wave (Tseng et al 2019), or both, as well as other remote influences (e.g. stratosphere). Identifying those mechanisms will help us to understand farther the predictability of these extreme events.

Figure 25 shows the longitude-time evolution of Z500 in the Northern Hemisphere for the analysis and forecasts. By 31 January, the analysis exhibits a regime transition from PNA+ to PNA-, with an anticyclonic anomaly between the date line and 160W and a cyclonic anomaly downstream. The PNA- persisted for the subsequent 3 weeks. Rossby wave trains are seen emanating from this area around the first 10 days of February, with an anticyclone crest reaching Central Europe by the second half of February (solid blue line in the left panel). Figure 25 shows that the 28 January 2019 forecasts were able to persist correctly the PNA- regime and represented well the sequence of Rossby waves travelling into

the Extratropics during the first 10 days of February. In contrast, the forecasts initialized on 24 January exhibit a short-lived PNA- transition and reverting after a couple of days to a persistent PNA+. The coherent Rossby wave trains are not present in this forecast. The establishment of the high anomaly over Europe during 11-17 February is associated with the propagation of those waves. Rossby waves might have contributed to the persistence of the PNA- regime. Failing to correctly represent the Rossby wave propagation at the medium range prevented the 24 January 2019 forecast from predicting realistic anomalies at the extended-range (week 3.5). Further analysis is in progress to establish the reasons for the loss of predictability between these two forecasts, initiated just 4 days apart.

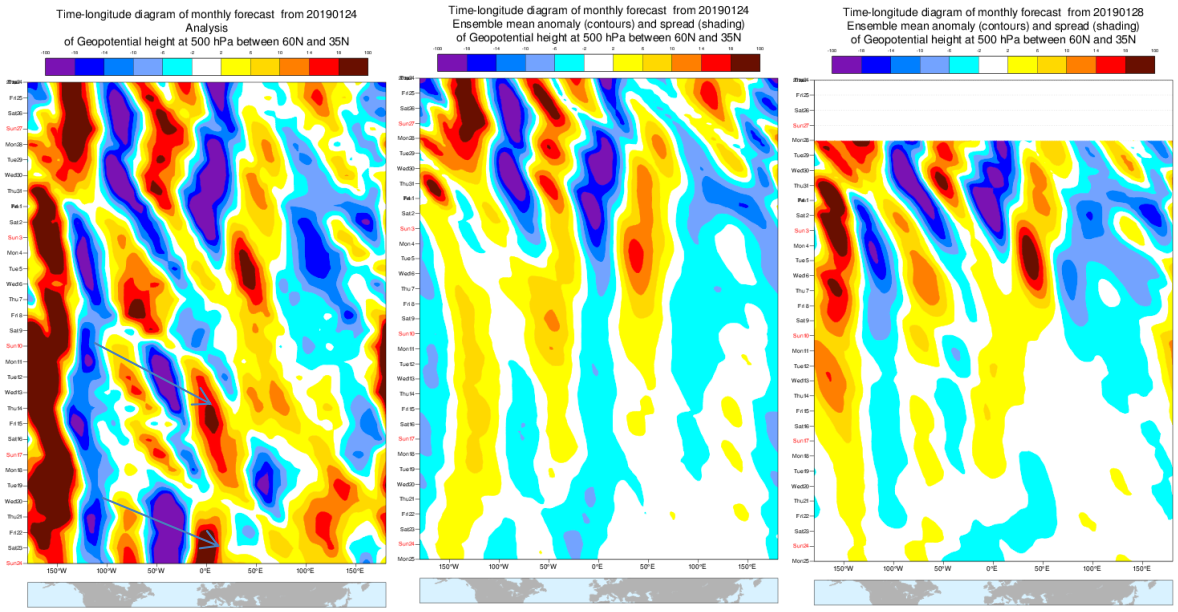

Figure 25: Time-longitude diagram displaying daily evolution of geopotential height anomalies at 500 hPa averaged between 60N and 35N. a) Analysis b) and c) ensemble mean for 24/01/2019 and 28/01/2019 forecasts. By 31 January, there analysis exhibits a transition from PNA+ to PNA- which persists for the subsequent 3 weeks. Rossby waves trains are seen emanating from this area on 11 February, with an anticyclone crest reaching Central Europe by mid-February (solid blue line in the left panel). The forecasts initialized on 24 January are unable to locate the phase of the PNA, producing a PNA+. The forecasts initialized on 28 January are able to position correctly the PNA phase, although in the ensemble mean it is weaker and less persistent. The ensemble mean gradually loses amplitude as the model predictability decreases and the ensemble decorrelates spatially.

6 Future directions

6.1 Horizontal atmospheric resolution

Since 2002, the atmospheric resolution of the extended-range forecasts has increased from Tl159 (about 110 km) to Tl255 (about 80 km), then to Tl319 (about 60 km) and eventually to Tco319 (about 36 km) in 2016. The next supercomputer in Bologna will bring an opportunity to increase the horizontal resolution even further. It is unlikely that there will be enough computing resources to increase the resolution of LegB (after day 15) beyond Tco639, which is the current resolution of LegA (before day 15). In order to assess the impact of the horizontal atmospheric resolution on extended-range forecast skill scores, 5-member ensemble re-forecasts have been run for 46 days at Tco639 starting on the 1st of each month from 1989 to 2016. The scorecard between this experiment and the control experiment

(Tco319 resolution from day 0) suggests a modest positive impact in week 1 (day 5-11), but no statistically significant positive impact after week 1 (Figure 26). The impacts on MJO, SSW, NAO forecast skill scores are also not statistically significant. Therefore, a future increase of horizontal resolution up to Tco639 is not expected to lead to significant improvements in extended-range forecasts.

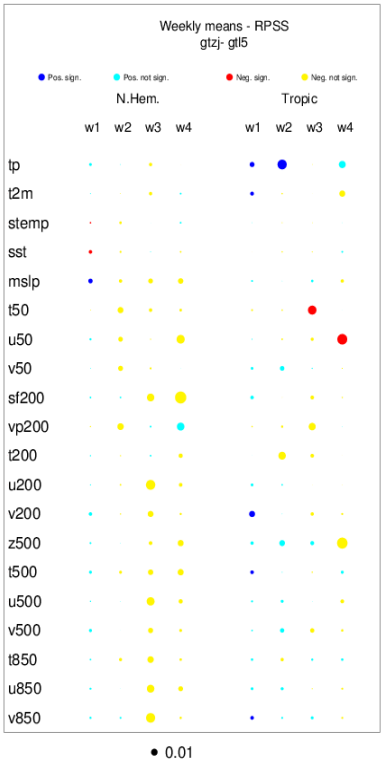

Figure 26: Scorecards of the difference of RPSS between the re-forecasts run at Tco639 (gtzj) and Tco319 (gtl5). The blue (red) colours indicate a positive (negative) impact when increasing resolution. Dark blue or red indicate that the difference is statistically significant using a 10,000 bootstrap resampling technique.

In order to test higher resolutions, a COLA-ECMWF collaborative project on high-resolution seasonal and sub-seasonal predictions supported by NCAR Accelerated Scientific Discovery programme, called METIS (follow up of projects Athena and Minerva) was launched in 2016. The coupled ensemble re-forecasts were run on the NCAR Cheyenne HPC. The ocean resolution was the same as currently in operations (¼ degree NEMO), but IFS cycle 43R1 was run with three different atmospheric resolutions: Tco199, Tco639 and Tco1279 with 91 vertical levels. Results suggest that there is no significant impact on the extended-range forecast skill scores when increasing the resolution to Tco1279 for 2-metre temperature, precipitation and 500 hPa geopotential height. As an example, Figure 27 shows the difference of RPSS of 500 hPa geopotential height between the high-resolution runs (Tco1279) and the low-resolution runs (Tco199). The difference is not statistically significant. The impact of increased horizontal resolution on MJO skill, amplitude and teleconnections is also not statistically significant. This is not to say that horizontal resolution does not play a role but suggests that it is not the primary source of errors at this forecast range. However, more in depth work is needed to analyse these runs.

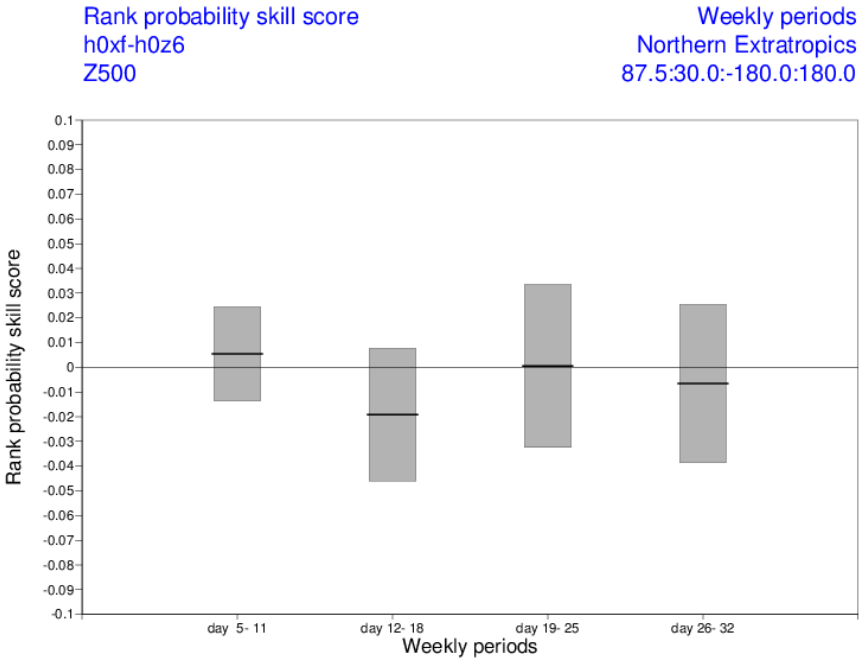

Figure 27: Difference of RPSS between the Tco1279 (h0xf) and Tco199 (h0z6) experiments from the METIS project based on 60 cases (1986-2015 November and May start dates). Negative (positive) values indicate worse (better) scores at Tco1279. The vertical bars indicate the 5% level of confidence.

6.2 Vertical resolution

The current vertical resolution of the Ensemble system is L91. The impact of increasing the vertical resolution to 137 vertical levels, as in HRES integrations, has been assessed by running a series of 5-member re-forecasts starting on the 1st of each month from ERA5 for the period 2000-2017, using cycle 46R1 with 91 vertical levels (control) and 137 vertical levels (EXP137). According to Figure 28, the cold bias in the tropical stratosphere is strongly reduced with 137 vertical levels, and the wind bias in the lower stratosphere is also reduced by up to 1 m/s. Therefore, increasing the number of vertical levels from 91 to 137 levels significantly improves the model climatology, as shown in Polichtchouk et al (2019). The impact on the forecast skill scores is also positive, with improved forecast skill in the Northern Extratropics in week 1, and in the Tropics up to week 4 (Figure 29). The impact on the MJO ensemble mean bivariate correlations is neutral. Interestingly, the 137 vertical level integrations produce significantly stronger MJOs (Figure 30), reducing the current bias in the extended-range forecasting system (too weak MJOs). Since the difference of tropical wind climatology in the lower stratosphere between EXP137 and Control is similar to the observed difference between an easterly and a westerly QBO, this result could be related to recent findings that the MJO is stronger during the easterly phase than during the westerly phase of the QBO (Liu et al, 2014; Yoo and Son, 2016).

Additional experiments have been performed to assess the impact of a further vertical resolution increase to 157 levels. Although the experiment with 157 levels produces a slightly warmer tropical stratosphere than EXP137, the impact on the forecast skill scores compared to the 137 levels experiment is not statistically significant (not shown). Preliminary experiments with a new sponge layer (higher than in the current system) and with quintic vertical interpolation (Polichtchouk et al, 2019) did not produce significant impact on the extended-range forecast skill scores, despite producing a slightly warmer tropical stratosphere. The impact on SSWs was also not statistically significant.

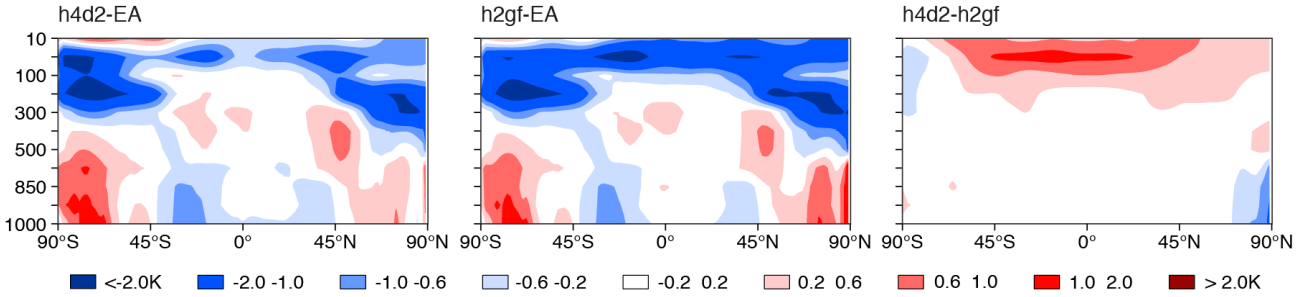

Figure 28: Zonal mean temperature biases for the week 4 forecasts (day 26-32), starting on 1 February, as a function of latitude (x-axis) and pressure level (y-axis) for EXP137 (left panel), Control (middle panel) and the difference between EXP137 and Control (right panel).

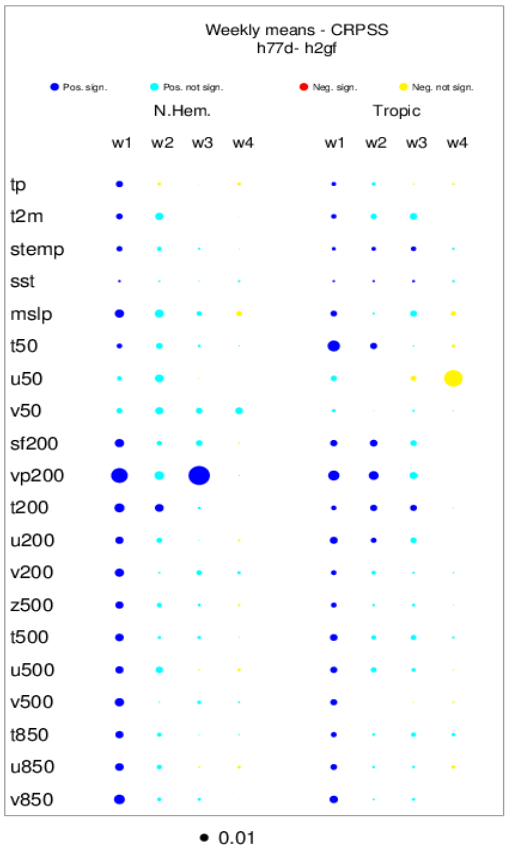

Figure 29: Scorecard of the difference of CRPSSF between the re-forecasts run with 137 vertical levels (h77d) and 91 levels (h2gf). The blue (red) colours indicate a positive (negative) impact when increasing vertical resolution. Dark blue or red indicate that the difference is statistically significant using a 10,000 bootstrap resampling technique.

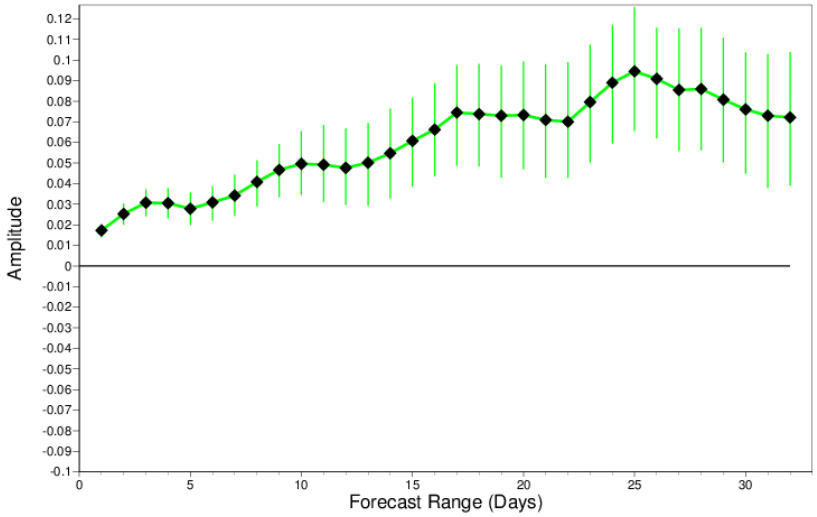

Figure 30: Difference of MJO mean amplitude between EXP137 (h4d2) and the Control experiment (h2gf) as a function of lead time. The vertical bars represent the 95% level of confidence using a 10,000 bootstrap resampling technique. The black diamonds indicate the days when the difference is statistically significant.

6.3 Ensemble strategy

This section discusses 3 different aspects of the ensemble strategy for extended-range forecasts:

- Frequency and ensemble size of the real-time forecasts
- Frequency and ensemble size of re-forecasts
- Variational resolution (VarEPS)

6.3.1 Extended-range real-time forecast strategy

Operational extended-range forecasts are produced twice a week, every Monday and Thursday from an extension of the 51-member ensemble to 46 days. Other centres (e.g. Environment Canada) adopt a similar strategy of large ensembles produced on a weekly or bi-weekly frequency. This ensemble generation is referred to as “burst sampling”. However, other centres (e.g. UKMO, NCEP, CMA) adopt a different strategy: the extended-range forecasts are produced daily, but with a small ensemble size (e.g. 4 members for UKMO). In order to inflate the ensemble size, real-time forecasts are produced by combining all the forecasts produced over a window of several days (7 in the case of UKMO). This produces a “lagged ensemble” in which ensemble members have different lead times. So far there has been no comparison between both ensemble generation strategies for extended-range forecasts. This section will investigate the advantages and disadvantages of running daily lagged ensemble extended-range forecasts at ECMWF.

A main advantage of burst sampling is that, for the same ensemble size (after lagging when applied), it should be more accurate and skilful than a lag ensemble on the burst start dates (Mondays and Thursdays). In addition, the forecast calibration and product generation is much simpler with burst sampling since all the members have the same lead-time. On the other hand, a lagged ensemble is likely to provide more skilful forecasts the other days of the week. The day to day evolution of the forecast products is also likely to be smoother, with less chances of flip-flop behaviour. We are interested in answering the following two questions, if the current 51-member ensemble run twice a week is replaced by an extended-range ensemble run every day with an ensemble size Ne:

- What is the minimum value of Ne so that there is a lagged ensemble forecast (Nd forecast days combined) which is at least as skilful as the current system on Mondays and Thursdays?
- For a given value of Ne, what is the optimal number Nd of forecast days to combine? Greater values of Nd produce larger lagged ensemble size, but also reduce the accuracy of the forecasts by adding ensemble members with older start dates.

In order to answer these two questions with the ECMWF model, a 15-member ensemble has been run for 32 days starting on 1 February, May, August and November 1989 to 2016. The resolution is Tco319 and the model cycle is CY45R1. These integrations represent the control burst sampling experiment. Additional 15-member ensembles starting 1, 2, 3 and 4 days before each start date of the Control experiment have been generated to allow the production of lagged ensembles with a window of up to 5 days. Since it was too expensive to run 51 member ensembles and since the answer to the 2 questions above is likely to depend on the burst sampling ensemble size (large burst ensembles are likely to be more difficult to beat than small burst ensembles), the values of the CRPS computed from the weekly mean anomalies of these experiments have been extrapolated to the CRPS of a larger ensemble using a formula from Ferro et a. (2018). This formula extrapolates the value of the CRPS of ensemble size M from the CRPS of an ensemble size m:

$$C_M = \frac{m(M+1)}{M(m+1)} C_m$$

This formula which, in principle, is valid only for reliable ensemble forecasts, has been tested on subsets of the 15-member ensemble giving reasonable approximation of the CRPS. Using this formula, we can compute the difference of CRPSS applied to weekly means and averaged over the 20 variables of the scorecard for each weekly lead time between a lag ensemble (with Ne ensemble members and a window of Nd days) and the 51-member burst sample. Figure 31 shows the difference for the Northern Extratropics as a function of the ensemble size Ne per day on the x-axis and the window size Nd on the y-axis for the weeks spanned by days 5-11, 12-18, 19-25 and 26-32. For the burst sample, the CRPS corresponds to the skill on the day the ensemble forecast is produced (Mondays or Thursdays in our current system). For other days of the week, the difference of skill scores would be much more to the advantage of the lagged ensemble.

According to Figure 31, the benefit of a lagged ensemble increases with lead time. For week 1, there is no advantage in replacing the current burst ensembles with a lagged ensemble for the Northern Extratropics when the lagged window is larger than 24 hours. Including ensemble members which were produced more than 24 hours earlier degrades the forecast skill at this time range even if the ensemble size is strongly inflated. However, this time range is included in legA and is already produced twice daily so it is not affected by running legB on daily basis rather than bi-weekly basis. For week 2 (corresponding here to days 12-18) forecasts, which cannot currently be produced daily, there are indeed lagged ensemble configurations which would produce more skilful forecasts than the current system on Mondays and Thursdays (pink area in the top right panel of Figure 31). To be at least as skilful as the current system, the minimum Ne should be 20 per day, and the optimal scores would be obtained for Nd=3, by combining the forecasts started on days 0, -1 and -2 (3-day window). For weeks 3 and 4, the minimum required ensemble size of the lagged ensemble gets smaller (Ne=14 members for week 3 and Ne=13 members for week 4), and the optimal window size of the lagged ensemble increases (Nd=4-day lag). It is interesting that at the longer time range (weeks 3 or 4), the pattern follows closely the curve of equal ensemble size (black line in Figure 31), suggesting that at this time range the ensemble size is more important than the degradation in the quality of the initial conditions, which is the opposite to the finding for week 1. Week 2 lies in between.

Over the Tropics (Figure 32) the benefit of a lagged ensemble is larger in weeks 1 and 2 than in the Extratropics. This figure shows that better day 5-11 weekly averaged forecasts in the Tropics could be

produced in the current system from a 1-day window lag ensemble. This result also suggests that more than 51 ensemble members are needed to predict the Tropics at 1-week lead time. For weeks 3 and 4, there are less differences between the Tropics and Northern Extratropics, although the optimal lag ensemble window is slightly longer in the Tropics.

Figure 33 summarizes these results over the Northern Extratropics and the Tropics, showing the minimum ensemble size per day and the optimal lagged window size for a lagged ensemble to be as skilful as the current Monday and Thursday extended-range forecasts for weeks 1 to 4. This figure shows that:

1. A lagged ensemble is more beneficial in the Tropics than in the Northern Extratropics particularly for shorter lead times (weeks 1 and 2).
2. The minimum daily ensemble size for legB (weeks 2 and beyond) when using lag ensemble is Ne=20, with an optimal number of lag days Nd=3.

The minimum configuration described in 2) would produce extended-range forecasts as skilful as the current system on Mondays and Thursdays, but better extended-range forecasts the other days of the week. This would be more expensive than the current system (140 members instead of 102 per week), but this would affect only the cost of legB in real-time forecast. Ideally, running legB with 51 members daily would give the flexibility of producing burst ensembles every day as we are currently doing twice a week, as well as giving the possibility of producing lag ensembles which would improve the forecast skill (CRPSS) for weeks 2 to 4. This would lead to an improvement of about 1% in the Tropics and northern Extratropics on Mondays and Thursdays in weeks 3 and 4 and, on average, by about 4% (8%) in the northern Extratropics (Tropics) if we consider all the days of the week, compared to the current operational system. This option would increase the cost of legB in real-time forecasts by a factor 3.4, but would increase the total cost of legB (real-time forecasts + re-forecasts) by only 47%, since the current cost of re-forecasts is much higher than the cost of real-time forecasts. Since legB represents about 20% of the cost of the whole ENS operational forecasts, such system would increase the cost of ENS by about 10%. It would benefit the users by allowing the production of new extended-range forecasts every day that would be more skilful than the current system. For users who need extended-range forecasts only once a week, this would also provide more flexibility on the choice of the forecast day.

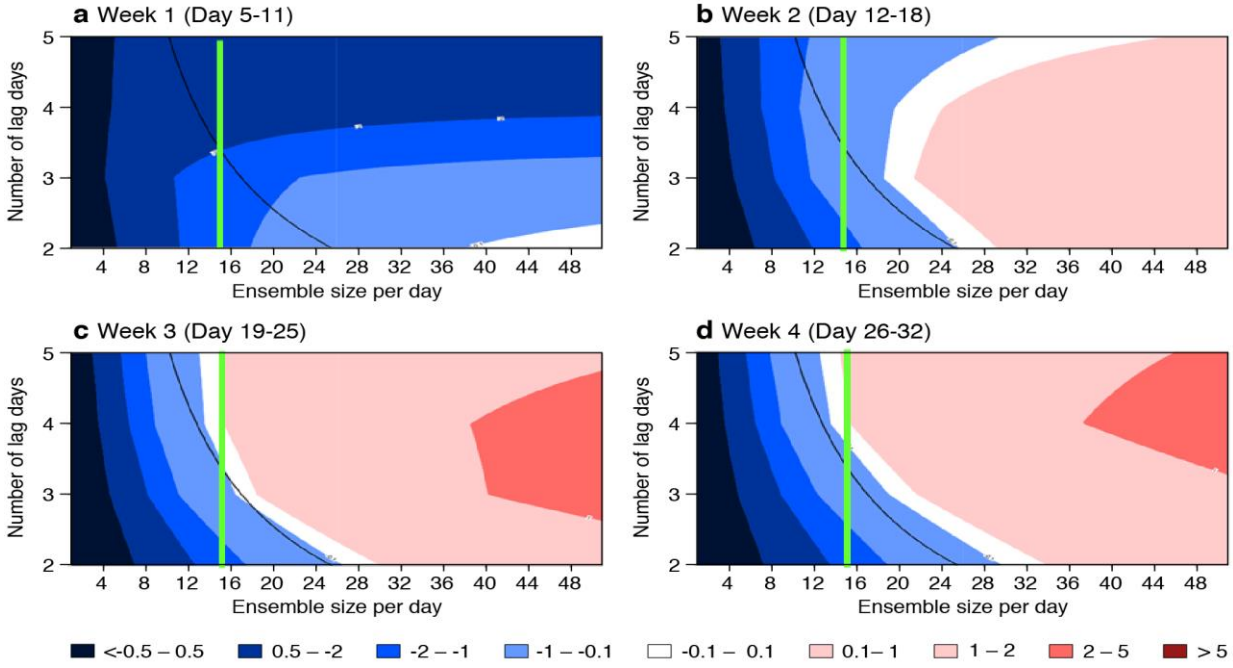

Figure 31: Difference of CRPSS between a lagged ensemble and a 51-member burst ensemble as a function of the daily ensemble size on the x-axis (from 1 to 51) and the number of forecast days combined on the y-axis (from 2 to 5). Blue (red) colours indicate that the burst ensemble is more (less) skilful than the lagged ensemble on the days the burst sample is produced. The difference is expressed as a percentage of improvement or degradation relative to the burst ensemble CRPSS. The black curve represents the line of equal ensemble size of the burst and lagged ensembles, while the green line indicates the number of ensemble members per day which would generate the same cost as the current system (51 members twice a week vs 14 members every day).

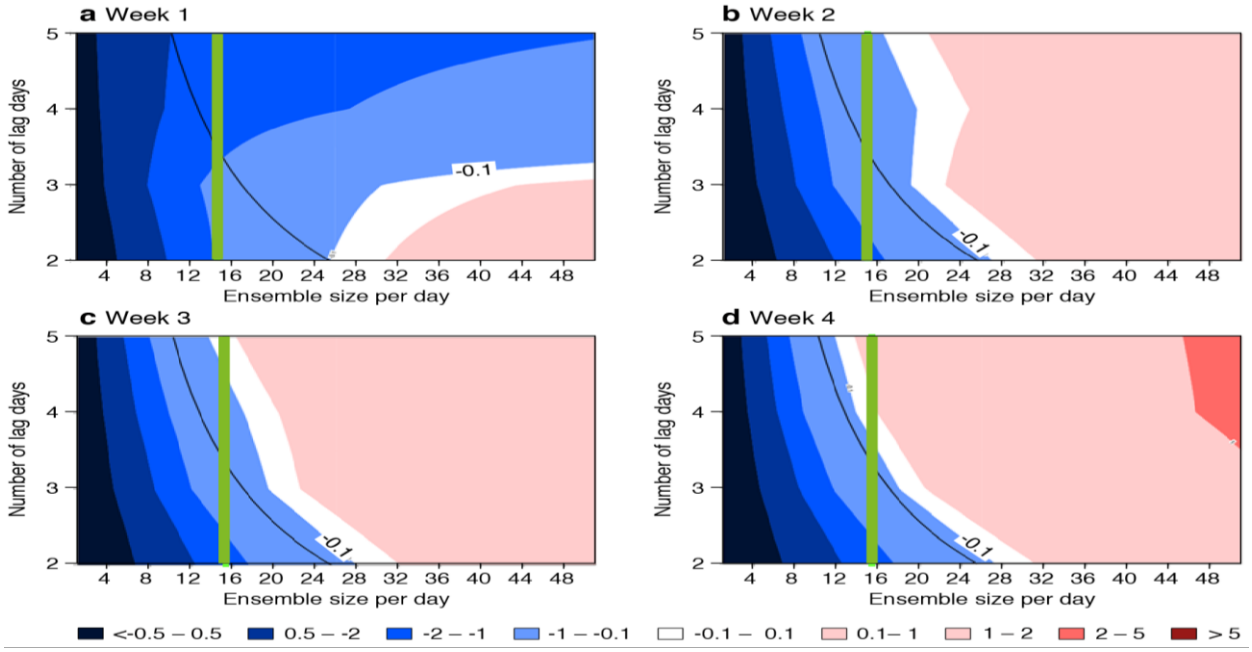

Figure 32: Same as Figure 32 but for the Tropics

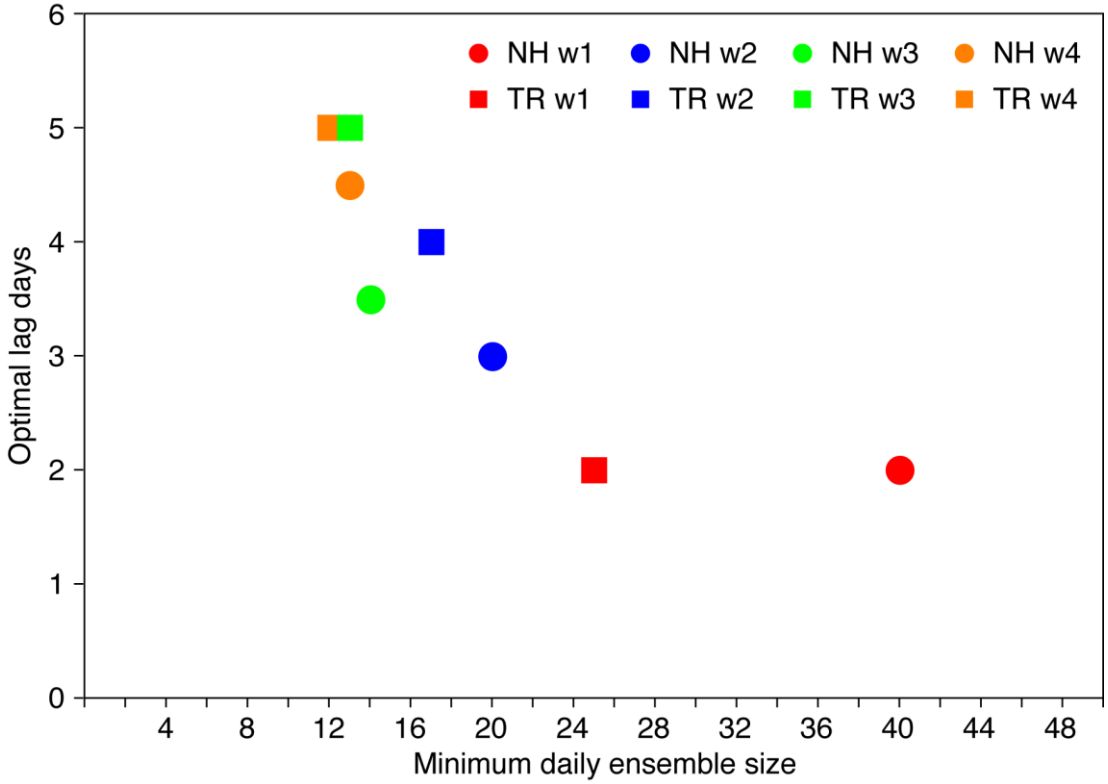

Figure 33: The minimum daily ensemble size (x-axis) and optimal number of lag days (y-axis) for weeks 1 to 4 over the Northern Extratropics (circles) and Tropics (squares) required for a lagged ensemble to be as skilful as the currently used burst ensembles for the days the burst ensembles are produced.

6.3.2 Variational resolution (VarEPS)

Another possible change, independent of the decision of running the extended-range forecasts daily (previous section), is to suppress the change of atmospheric horizontal resolution at day 15. In the current system, the first 15 days (LegA) of the extended-range forecasts are run at a higher resolution (Tco639) than after day 15 (LegB) which is run at Tco319. This change of resolution makes the generation of products difficult particularly for accumulated variables such as precipitation. According to Figure 34, there is no statistically significant difference in the forecast skill of weeks 2, 3 and 4 between an experiment with the current operational configuration and an experiment at LegB resolution from day 0, although the first week benefits from the higher resolution. Therefore, a possible future configuration for the next HPC could be to produce extended-range forecasts from a low-resolution model (higher or equal to Tco319). The horizontal resolution of this system would depend on the configuration of the next HPC. This new system, which, in the following, we refer to as Low Resolution ensemble (L-ENS) would be separated from the medium-range higher-resolution ensemble. Although this would look like a reversal of the move in 2008, when medium and extended-range forecasts were merged into a single system with variable resolution (VarEPS), the future developments of both systems would remain strongly coupled. In addition, this new configuration would be more versatile and allow the production of medium-range dual resolution ensembles, which was not considered in 2008 when VarEPS was designed. There are several advantages from this design:

1. It would remove the issues linked to the change of resolution at day 15.

2. It offers more flexibility to reconfigure the medium range (MR) forecast and reforecast production. The medium-range (MR) reforecast are needed for the EFI. The EFI will benefit by having more frequent sampling, but it does not need so many ensemble members.
3. This new system could be used to produce daily 51-member low resolution medium-range forecasts up to day 15 which could be added to the 51-member higher resolution medium-range forecasts to produce a 102-member ensemble of medium-range forecasts and therefore be part of the medium-range ensemble forecasting system.

However, a downside would be an increased computer cost for running an additional LegA, even if it is at a lower resolution.

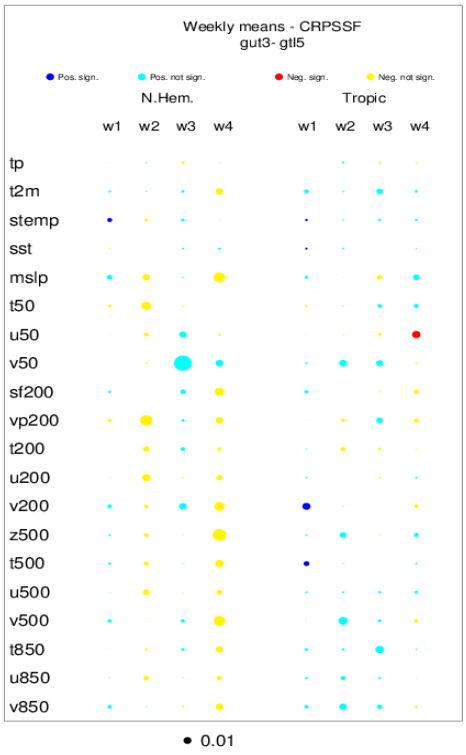

Figure 34: Scorecards of the difference of CRPSSF between the re-forecasts run at Tco639 up to day 15 and Tco319 up to day 46 (gut3) and Tco319 from day 0 to 46 (gtl5). The blue (red) colours indicate a positive (negative) impact when increasing resolution in LegA. Dark blue or red indicate that the difference is statistically significant using a 10,000 bootstrap resampling technique.

6.3.3 Re-forecast ensemble strategy

If extended-range real-time forecasts are run daily, then the current configuration of the re-forecasts will need to be revisited. There will be no longer a justification for running them twice weekly. Instead a higher frequency, daily or every two days, depending on the computing resources, would be very helpful for skill assessment or analysing past cases studies. However, a minimum ensemble size of 5 would be needed.

Extended-range re-forecasts are used for calibration of the real-time forecasts as well as for skill assessment. The requirements for calibration and skill assessment are not the same. The re-forecast ensemble size is less important for calibration than for skill assessment. Calibration would benefit from more frequent and longer re-forecast periods rather than large ensembles. However, skill assessment is more accurate with large ensembles. Some scores, like the CRPSS, can be adjusted to take into account the impact of ensemble size, but assessing the skill of extreme events requires larger ensemble size. A possible re-forecast configuration to satisfy both requirements could be the following:

- 25 members on the 1st of each month (same as for seasonal, except that it is produced routinely) for skill assessment
- 5-member ensemble every other day for calibration and for some skill assessment when fair score is used. Additional start dates would also be useful also for predictability studies

The cost of this configuration would be almost the same as in the current operational system (about 100 members per re-forecast year and month).

6.3.4 Summary and cost benefits of proposed changes

The sections above have discussed several possible changes to the configuration of the extended-range forecasting system. Figure 35 shows a comparison of the benefit (improved CRPSSF skill for weeks 2, 3 and 4) and computing cost of each of these options, in addition to the impact of ERA5 for comparison.

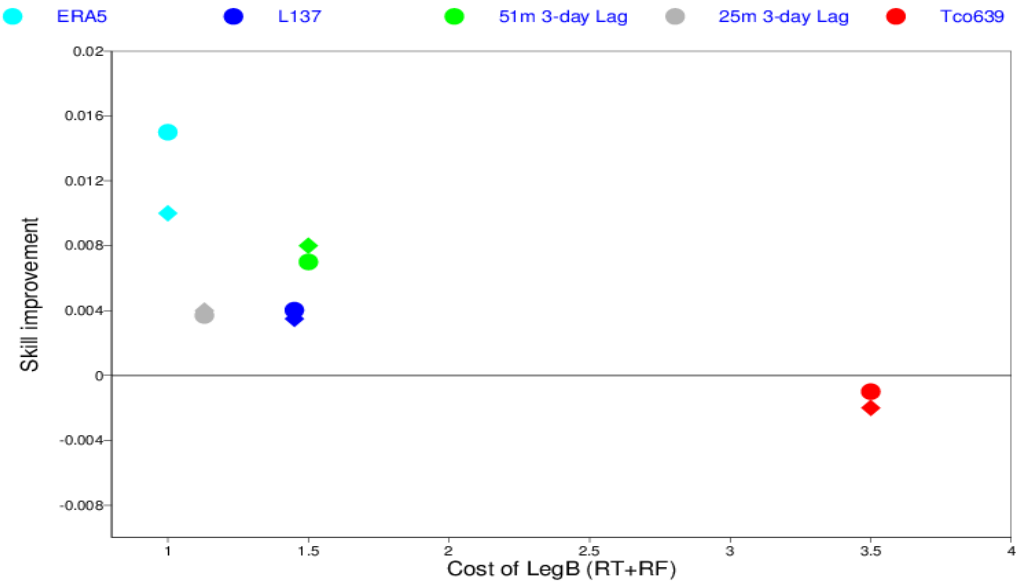

Figure 35: Cost benefit of the various changes proposed for the next configuration of the extended-range forecasting system. The x-axis indicates the increased CPU time for each individual change as a ratio with the current cost of LegB (real time and re-forecasts). The y-axis shows the expected skill improvement (CRPSSF averaged over 20 variables and weeks 2, 3 and 4) relative to the current operational system. Circles are for the northern Extratropics and diamonds for the Tropics. Cyan symbols represent the impact of ERA5 initialization on re-forecasts skill scores, as a reference. Blue symbols show the impact of increased vertical resolution (137 vertical levels compared to 91 levels). Green (grey) symbols show the impact of running 51 (25) members daily with 3-day lag ensemble on the Mondays and Thursdays forecast skill scores. Red symbols show the impact of increasing the horizontal resolution of Leg B from Tco319 to Tco639.

Based on Figure 35, the following hierarchy between the various changes is proposed:

1. Run the real-time extended-range forecasts daily, instead of twice weekly, ideally with 51 ensemble members.
2. Run a small ensemble of re-forecasts with a daily or bi-daily frequency instead of twice weekly. Run a larger ensemble once a month for a more accurate skill assessment. This change could be done without increasing the computing cost.
3. Increase the vertical resolution to the same resolution as the High-Resolution system, while benefitting from a 30-40% model speed up through the use of single precision.
4. Run the extended-range forecasts separately, at low resolution from day 0.
5. Increase the atmospheric horizontal resolution.

All the changes discussed so-far in Section 6 would be, if agreed, implemented soon after the new HPC becomes operational in Bologna. The sections below discuss longer-term changes.

6.4 Interactive aerosols

Extended-range forecasts have benefitted from increased model complexity, such as the introduction of an interactive sea ice model and ocean-atmosphere coupling from day 0, over the past decade. The next direction to explore is the inclusion of atmospheric chemistry, such as ozone and aerosols. In particular, the radiative effects associated with aerosols have been shown to have visible impact in model climate (Rodwell and Jung, 2008; Bozzo et al, 2017). The climatology described by Tegen et al (1997) has been used for several years with satisfying results (Tompkins et al, 2005; Rodwell and Jung, 2008). Recently, updated aerosol climatologies based on the Copernicus Atmosphere Monitoring Service (CAMS) interim reanalysis (hereafter CAMSira; Flemming et al, 2017) have been used to replace the Tegen et al (1997) climatology (Bozzo et al. 2017). These changes had positive impacts on the extended-range forecast skill scores, suggesting that the representation of aerosols can play a significant role in S2S prediction.

In the current Ensemble system, the representation of aerosols is still based on climatology. However, the observed aerosol distribution can display large inter-annual and intra-seasonal variability. For instance, the MJO plays an important role in aerosol variability as first pointed out by Tian et al (2008) and further discussed by Tian et al (2011) and Guo et al (2013). The authors analysed several years of aerosol optical depth (AOD) data from the MODIS retrievals and concluded that some of the modes of variability in AOD was indeed correlated to the time scales of the MJO. The MJO-related intra-seasonal variance accounts for about 25% of the total aerosol optical thickness variance over the tropical Atlantic (Tian et al, 2011), primarily through its influence on the Atlantic low-level zonal winds. This aerosol variability is of course not present in the current aerosol climatology used for the extended-range forecasts. Several experiments, performed under the funding of the DACCIWA project, have been performed to assess the impact of having time-varying radiatively interactive aerosols on the extended-range forecasts skill scores. In these experiments, only the direct effect of aerosols is accounted for. The aerosol direct effect consists of the sum of two phenomena: scattering/absorption of incoming solar radiation and absorption/emission of longwave radiation. The radiative impact of aerosols is very dependent on their vertical distribution, chemical composition, and surrounding environment. These experiments are also an opportunity to explore the predictability of aerosols which could be useful for some applications such as air quality and health (e.g. meningitis).

Four experiments were run with cycle 43R1 to assess the aerosol impacts: one control integration with the Tegen (1997) climatology in which all settings were similar to the extended-range operational setup,

but run at lower horizontal resolution (Tl255); a second control run with the CAMS/Bozzo climatology at the same reduced resolution (CONTROL2); an interactive prognostic aerosol run in which the prognostic aerosols are initialized using the time-varying CAMSira (PROG1); and a second interactive aerosol run in which the prognostic aerosols are initialized using a fixed climatology, similar to the one used in CONTROL2, based on a CAMS experiment without any data assimilation (PROG2; J. Flemming, 2016 personal communication). The different initializations allow us to understand the sensitivity of the interactive aerosol runs to the aerosol initial conditions. The experiments PROG1 and PROG2 above are different from the standard NWP integrations in that the radiation interacts directly with the prognostic aerosols, instead of using the Tegen (1997) or the CAMS/Bozzo climatologies. All simulations were conducted with 91 vertical levels. Prescribed emissions for the anthropogenic species over the years of interest (2003–2015) were used in experiments PROG1 and PROG2. Aerosols are forecast within the global system by a bulk–bin scheme (Morcrette et al, 2009), based on earlier work by Reddy et al (2005) and Boucher et al (2002), that includes five species: dust, sea salt, black carbon, organic carbon, and sulphates.

Figure 36 shows the ranked probability skill score (RPSS) for the dust optical depth forecasts from the PROG1 and PROG2 integrations for the Tropics, verified against CAMSira. Persistence, which is usually more skilful than climatology at the extended time-range, is also shown for comparison. Both re-forecast experiments have higher RPSS than persistence for dust aerosol anomalies. Although both PROG1 and PROG2 have the same prescribed aerosol emissions, PROG1, which was initialized by CAMSira, scores the highest, highlighting the importance of the aerosol initial conditions. This result also hints at the potential added value of extended-range predictions for health-related applications. Dust has been linked in Northern Africa with outbreaks of meningitis. Having an even moderately skilful model prediction a month ahead could be useful for planning and preparation by the health authorities.

The impact on forecast skill scores of meteorological variables has been discussed in Benedetti and Vitart (2018). Results show the potential of interactive prognostic aerosols to improve model prediction at the monthly scales. Temperature and wind biases were reduced in both prognostic aerosol runs over several regions in the Tropics and the midlatitudes. When compared with CONTROL1 and CONTROL2, scorecards show positive impacts of the prognostic aerosols on several meteorological fields, including upper-level winds and lower-tropospheric temperature, particularly over the Northern Hemisphere (see Figures 4 and 5 in Benedetti and Vitart, 2018). The climatological initialization scored generally better than the CAMSira initialization in the first weeks possibly because the meteorological initial state, which was produced using climatological aerosols for radiation, is more in balance with the climatological initialialization than with the time-varying CAMSira initial conditions. These results suggest that having an interactive representation of aerosols may benefit extended-range prediction. They also show that the meteorological scores show sensitivity to the initialization of the aerosols, but more work is needed to understand this sensitivity. Running interactive aerosols currently increases the cost of the extended-range forecasts by about 50%. Further experimentations are coordinated by WGNE and the WWRP/WCRP S2S Project to better understand the impact of the individual species. This could help reduce the cost by having less species interactive and using climatology for the others.

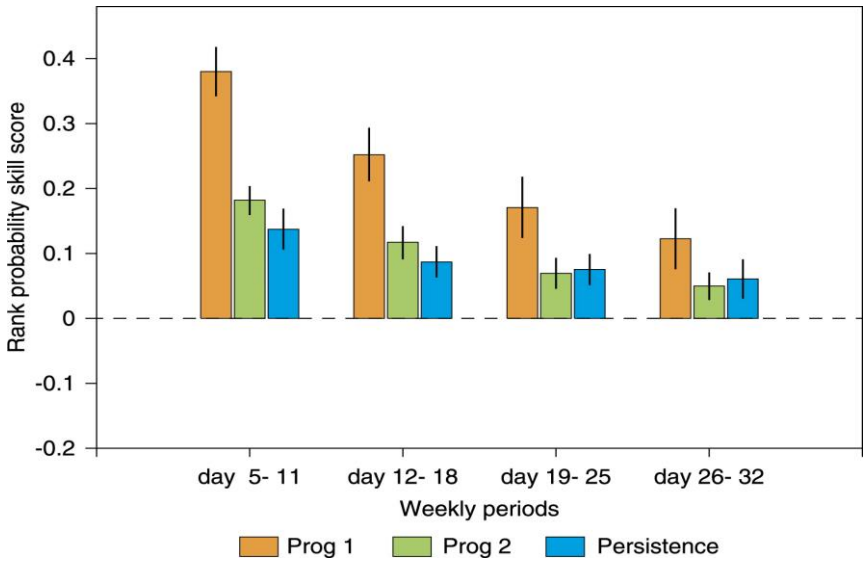

Figure 36: RPSS for experiments PROG1 (orange) and PROG2 (green) with respect to a persistence forecast (blue) of dust optical depth for the Tropics. CAMSira is used for the verification.

6.5 Initialization of the coupled system

Aside from the general realization that consistency between reforecast and real time forecast in the slow components of the Earth System (ocean, sea-ice and land) is required to avoid spurious anomalies, not much attention has been paid so far to the initialization of the coupled model for the extended-range forecasts. As part of the GLACE experiments (Koster et al, 2011), there have been targeted experiments aiming at quantifying the memory of the land initial conditions. However, the design of the experiments measured the impact of the interannual variability in the land conditions, rather than the impact of the observing system or assimilation methodology. The impact of atmospheric initial conditions from ERA5, visible at week 3, as reported in previous sections, is promising and prompts the need for the evaluation of major data assimilation developments using the extended-range forecasts. This is especially relevant in the context of the initialization of the Earth System, with foreseen developments in coupled data assimilation (CDA) and utilization of new observations in the years to come. Here we report on three specific aspects: i) the impact of coupled data assimilation on forecast of MJO and tropical cyclones; ii) the potential impact of observations of sea-ice thickness and iii) the impact of assimilating altimeter derived sea level anomalies.

6.5.1 Coupled data assimilation

CERA-SAT (Schepers et al, 2018) is a proof-of-concept coupled reanalysis with the full observing system available in the satellite era. It has been produced within the scope of the ERA-CLIM2. It uses the CERA coupled assimilation system (Laloyaux et al, 2016), which includes the atmosphere, ocean, land, waves and sea ice. It spans the period 2008-2016, and it uses an EDA of 10-ensemble members. CERA-SAT uses the IFS version Cy42R1 at a horizontal resolution of TL319 (~65 km) and 137 vertical levels. The ocean and sea ice components are based on NEMO v3.4 and LIM2 models, with the same version and resolution as in ORAS5 (e.g. approximately 25 km). The CERA assimilation system is based on a variational method with a common 24-hour window shared by the atmospheric, ocean and sea ice components. This is longer than the 12h standard assimilation window used for the atmosphere, but shorter than the window length used in ORAS5 (which is 5 days). The coupled model is introduced at

the outer-loop level, while the inner loop remains uncoupled. As in ORAS5, the air-sea interface in the ocean component is relaxed towards the high-resolution OSTIA sea-surface temperature analysis. The atmospheric model will see an SST analysis modulated and filtered by the ocean component, which implies errors as well. Schepers et al (2018) provide an evaluation of CERA-SAT from an atmospheric data assimilation perspective, by comparing CERA-SAT with an equivalent atmospheric reanalysis experiment in uncoupled mode, referred in what follows as CERA-SAT-Uncoupled.

Different combinations of atmospheric and ocean initial conditions from CERA-SAT, CERA-SAT-Uncoupled and ORAS5 have been used to evaluate the impact of coupled data assimilation (CDA) on the prediction of MJO and tropical cyclones. A set of extended-range forecasts were conducted to separate the impact of CDA in the atmosphere and in the ocean components. Experiments are a series of 46-day hindcast covering the period 2 May 2015 to 17 December 2016, with a 5-day interval (a total of 120 cases). Table 1 describes these experiments. Experiment CPL was initialized by the ocean and atmospheric components of CERA-SAT. In experiment UCP, the atmospheric component was initialized by CERA-SAT-Uncoupled, and the ocean component by ORAS5. Experiment HYBRID uses the atmosphere initial conditions from CERA-SAT, and the ocean initial conditions from ORAS5.

EXP-NAME	**ATM**	**OCN**	**ENS**
CPL	CERA-SAT	CERA-SAT	5
UCP	CERA-SAT Uncoupled	ORAS5	5
HYBRID	CERA-SAT	ORAS5	5

Table 2: main characteristics of the experiments to test the impact of coupled data assimilation in CERASAT on extended-range forecasts

The left panels of Figure 37 show the impact of the different initialization options on the forecast skill (RMSE) of the MJO, as measured by the Wheeler and Hendon bivariate index. The top panel shows the differences in RMSE between experiments CPL vs UCP, gauging the impact of coupled DA in both ocean and atmosphere. It shows that CPL has significantly degraded skill during the first 20 days of the forecast. The middle panel, showing the differences CPL-HYBRID measures the impact of CDA in the ocean, and it shows that the degradation is attributed to the ocean component. The impact of CDA in the atmospheric initial conditions, measured by comparing HYBRID versus UCP and shown in the lower panels, is positive between days 10 and 20, although the statistical significance is marginal. The anomaly correlation coefficient shows similar conclusions. Further inspection reveals that this impact is stronger when the MJO convection is around the Maritime Continent and Western Pacific.

To gain insight into the physical mechanisms related to this degradation, a moist static energy (MSE) budget analysis has been conducted. Previous studies indicate that MSE is a useful tool for diagnosing the MJO convection (DeMott et al, 2016). The MSE budget analysis has been applied to the different experiments. Results indicate that the underestimation of latent heat flux during forecast day 0-10 in experiments initialized with the ocean component of CDA is responsible for the MJO forecast degradation. Such underestimation of latent heat flux is associated with underestimation of SST anomalies in initial condition. This is illustrated in the right panels of Figure 37, which shows the differences in SST and latent heat between the experiments when the MJO active convective phase is over the Indian Ocean/ Western Pacific. The experiments initialized with the ocean component of CDA

exhibit colder SSTs and reduced latent heat flux. The reasons for this behaviour need to be explored further. Inspection of the ocean initial conditions show that the CDA produces deeper ocean mixed layers, which in warm pool regions tend to reduce the values of SST and the associated latent heat flux.

Interestingly, all the experiments develop a marked initialization shock during the first two days into the forecast, characterized by a rapid decline in the intensity of the MJO, which saturates at 90% of the observed intensity (not shown). This initialization shock is similar to the one present in the current operational system. This indicates that important imbalances between the initial conditions and the model states still exist even in the initial conditions produced by CDA.

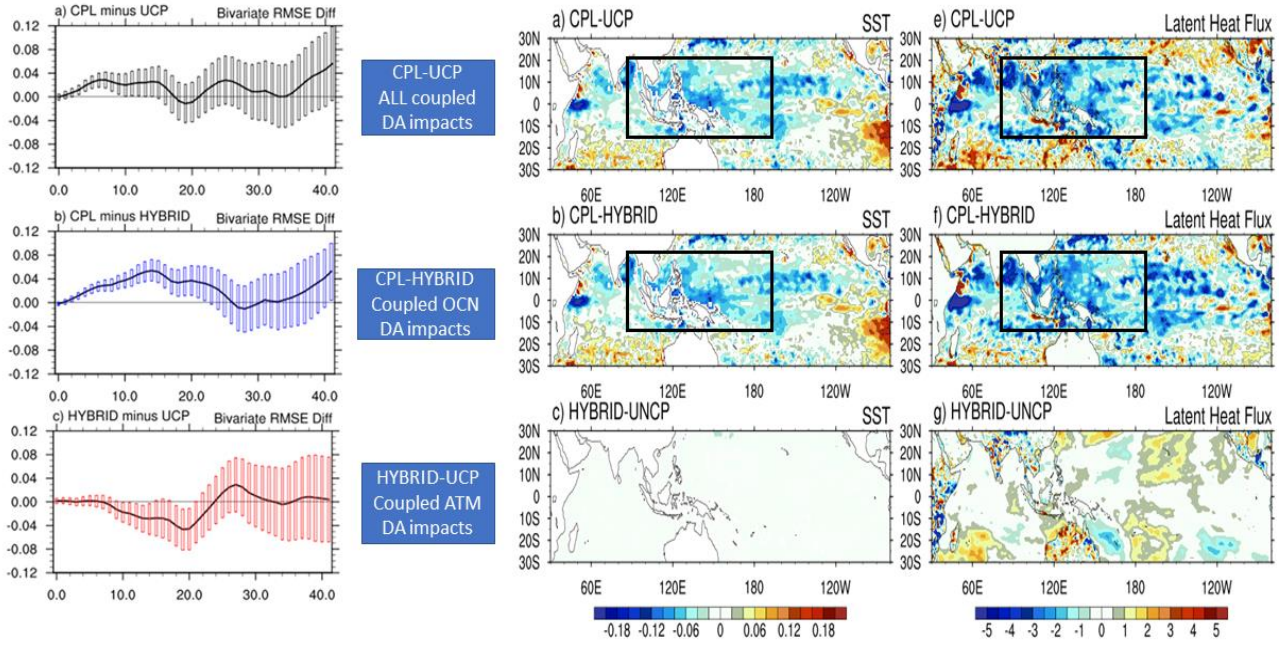

Figure 37: Impact of coupled data assimilation (CDA) on forecasts of the MJO. The first row shows the differences between experiments CPL v UCP, the middle row shows the impact of the ocean component of the CDA, (CPL-HYBRID) and the bottom row shows the impact of CDA in the atmospheric initial conditions. Left panels: Differences in forecast skill of the bivariate MJO index. Middle and right panels: difference in SST and latent heat flux anomalies corresponding to the active phase of the MJO. The ocean component of the coupled DA degrades the skill of the MJO in the first 20 days, due to underestimation of the latent heat flux forcing. From Yao et al. 2019, in preparation.

The experiments in Table 2 are further explored to evaluate the impact of CDA on the prediction of tropical cyclones (TC). There is little sensitivity to the TC track density among experiments. However, the coupled DA improves the TC intensity within forecast day 3-8 in the Eastern North Pacific (ENP) region. This improvement is mainly attributed to the atmospheric component of the CDA. An analysis of TC related variables suggests that the improved (stronger) TC intensity over this region is likely due to the differences in low-level relative humidity and air temperature in CDA as shown in Figure 38. According to Schepers et al 2018, the CDA improves the fit of these analysis variables to the observations in the Tropics.

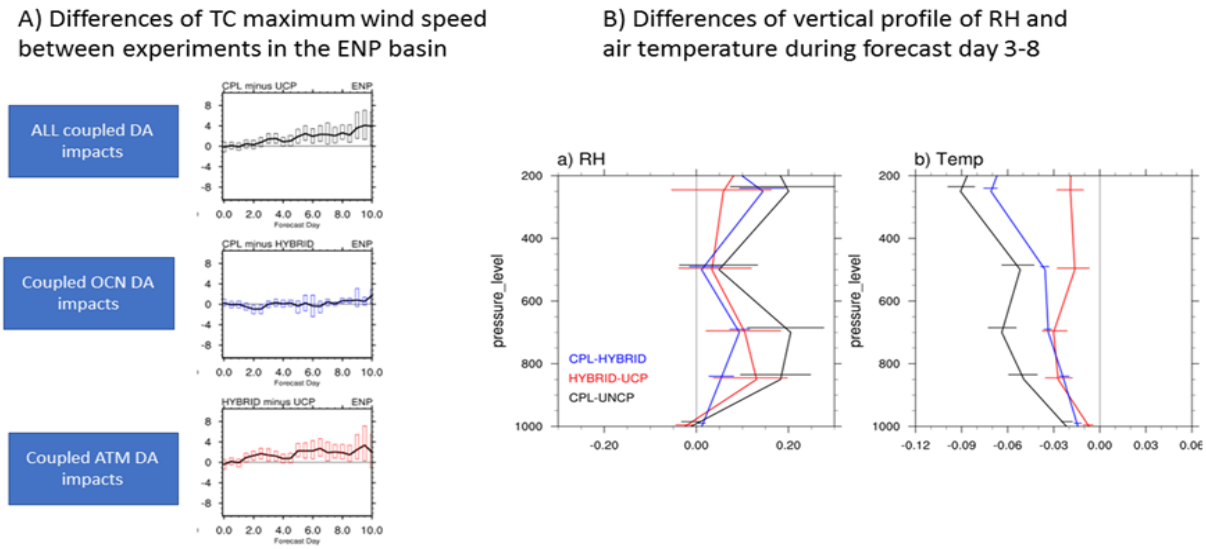

Figure 38: Impact of coupled data assimilation in forecast of tropical cyclones (TC) over the Eastern North Pacific. The atmospheric initial conditions produced with CDA increases the intensity of the TC as shown by the left panels, which show the difference in TC wind speed for CPL v UCP (top), CPL v HYBRID (middle) and HYBRID v UCP. The right panels show profiles of vertical differences in relatively humidity (RH) and temperature from the different experiments.

6.5.2 Impact of sea-ice thickness

As discussed in Section 3.4, ECMWF extended-range sea-ice forecasts are the best among the S2S models. As the initial sea-ice thickness is a good predictor for sea-ice cover changes in the extended-range, further improvements seem achievable if newly available observations of sea-ice thickness were to be used in the initialization and this particularly during the melt season.

Re-forecast experiments have been performed with Cy43R3 in a Tco369/ORCA1 resolution (Balan-Sarojini et al, 2019) with the Arctic sea-ice initial conditions obtained by constraining the sea-ice thickness to the CS2SMOS data set using nudging with a 10-day timescale. CS2SMOS provides weekly-mean sea-ice thickness for the winter months starting in 2011 by combining radar altimetry retrievals from CryoSat2 for thick ice with L-band radiometry retrievals for thin ice. All other ocean and sea-ice data assimilation settings, in particular sea-ice concentration assimilation using 3DVAR-FGAT, are the same as in the operational system OCEAN5. Re-forecast experiments start every month from January 2011 to December 2016.

Despite the shortness of the re-forecast period, we obtain significant improvements in forecasting the presence of sea ice during the melting season March to June. Figure 39a shows that the integrated ice edge error (IIEE) of weekly-mean sea-ice cover forecasts is reduced by up to 10% for longer lead times. At the same time, SST forecasts beyond the first week are significantly improved (Figure 39b). Interestingly, the skill differences are not significant for the first week, but keep increasing afterwards, suggesting that an improved initialization of sea-ice thickness clearly benefits extended-range more than medium-range forecasts.

Contributions to the integrated skill changes come mainly from regions where the seasonal sea ice edge resides, although improvements in the Hudson Bay, Labrador Sea, and the North Pacific dominate (Figure 40a). The forecast skill improvements for SST are mostly co-located with those for sea-ice cover (Figure 40b), because an improved prediction of the ice-free date allows an improved forecast of the

amount of seasonal warming of the SST. Work is ongoing to assess the impact of the improved se-ice cover prediction on the atmospheric circulation and skill scores.

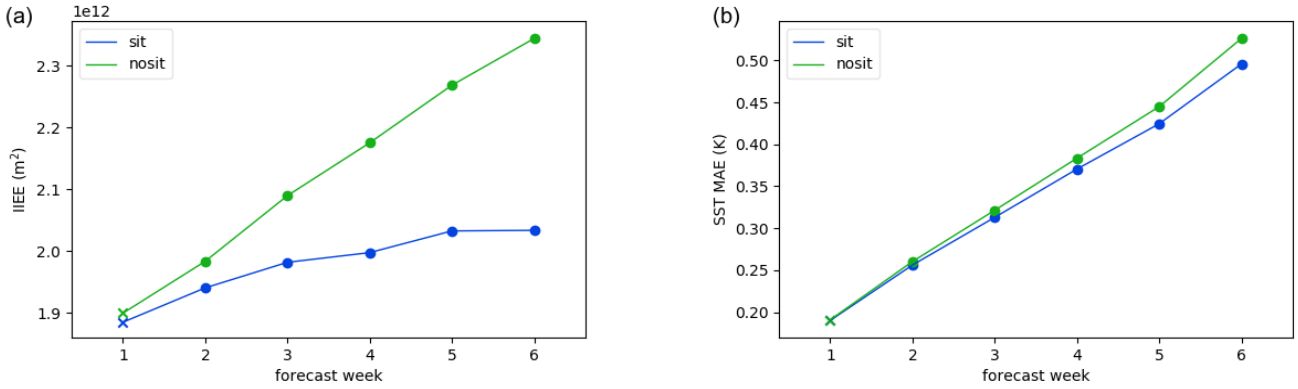

Figure 39: Spatially integrated forecast scores for ocean points north of 50N based on weekly means for the experiment with sea-ice thickness initialization (sit) and the reference experiment (nosit). Forecasts considered are started 1st of the months March-June 2011-2016 (24 cases) and have 25 ensemble members each. (a) shows the integrated ice edge error (Goessling et al, 2016) and (b) shows the mean absolute error of SST. Filled circles indicate significant differences according to DelSole and Tippett (2016), crosses otherwise

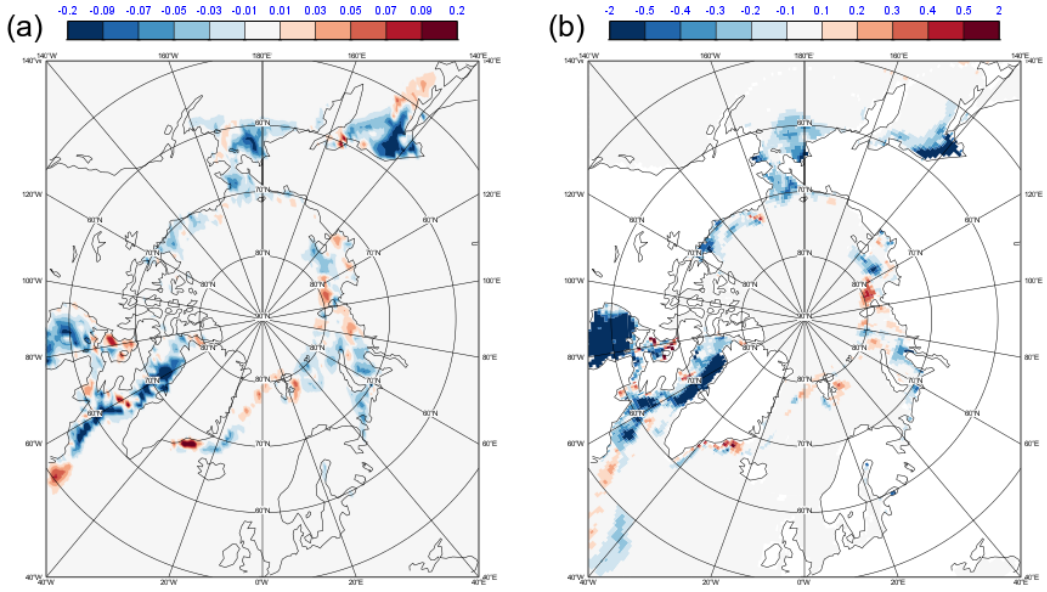

Figure 40: Change of forecast skill between the experiment with sea-ice thickness initialization and the reference for forecast week 6. The forecasts considered are the same as in Figure 39 (March-June 2011-2016, 25 ensemble members). (a) shows the change in Brier score for predicting the presence of sea ice with a 15% area fraction threshold, (b) shows the change in mean absolute error in K of SST forecasts.

6.5.3 Impact of altimeter-derived sea-level anomalies

The altimeter-derived sea level anomalies are a key component of the ocean observing system, and since 2006 (ORAS3) have been routinely assimilated for the production of ocean reanalyses that initialize the coupled forecast and reforecast, albeit with different methodology (Vidard et al, 2008, Balmaseda et al, 2013; Zuo et al, 2019). A few studies have demonstrated the value of this component of the observing system in ocean reanalyses and in seasonal forecast (Balmaseda and Anderson 2009; Zuo et al, 2015), but the value for the extended-range forecast has not yet been evaluated. This is a gap that the second phase of the S2S project attempts to bridge. As part of this program, a set of observing system experiments have been conducted at ECMWF, where specific elements of the ocean observing systems have been withdrawn in the production of ocean reanalyses which are later used to initialize extended-range forecast. Here we report on the results from one of those experiments, which assesses the value of altimeter data for the extended-range.

The procedure for the observing system experiment is as follows: in first instance, an ocean reanalysis equivalent to ORAS5 (Zuo et al, 2019), with identical model version, resolution, data assimilation and forcing, but without altimeter assimilation was conducted spanning the period 1993-2017. We call this 24-year ocean reanalysis ORAS5-NoAlti. In a second stage, two sets of extended-range reforecast experiments were produced, initialized by ORAS5 and by ORAS5-NoAlti (experiment gugh and gvur respectively). The experiments use the same coupled model configuration (cycle 45R1, Tco319-91 levels for the atmosphere, and ORCA025 for the ocean), and consist of 288 initial dates (24 years, forecast initialized every month), with 5 ensemble members for each date. The forecast skill from these two experiments was evaluated. The scorecards showed a statistically significant but very small impact for u850, v850 and mslp in the Tropics. The impact was more noticeable in the forecast of tropical cyclones (Figure 41). Figure 41 shows that by including altimeter information in the ocean initial conditions the skill of TC improves over most basins, as measured by the correlation in Accumulated Cyclonic Energy (ACE). The difference of score between both experiments is statistically significant over the western North and Central Pacific. This result is consistent with previous empirical studies showing that anomalies in the upper ocean (~50-100m) thermal structure precede the occurrence of ACE anomalies (Scoccimarro et al, 2018).

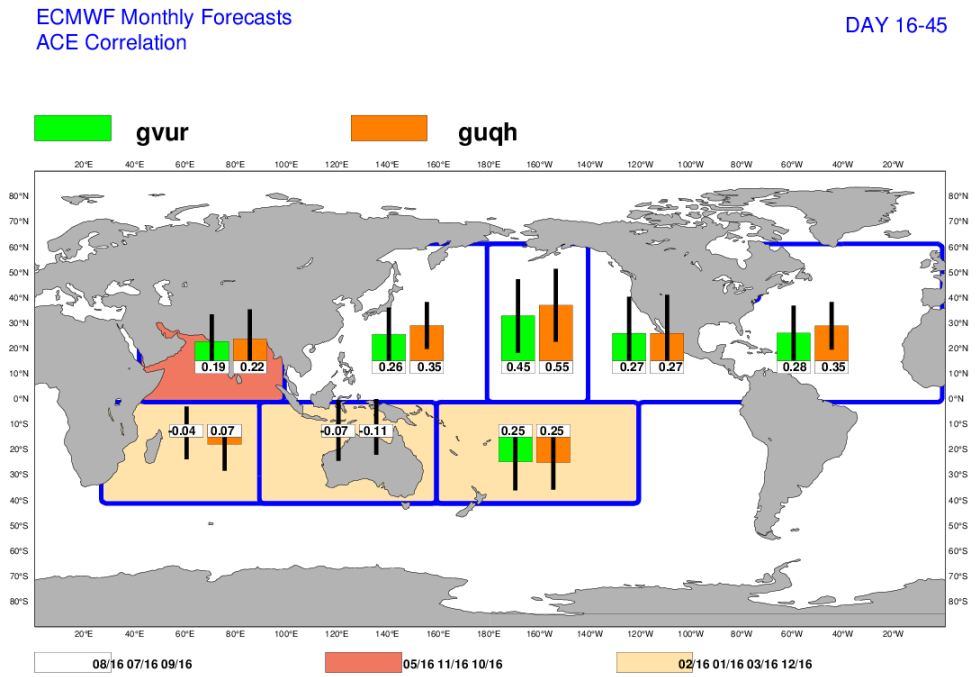

Figure 41: Anomaly correlation skill of extended-range forecast of the Accumulated Cyclonic Energy from experiments initialized by ocean reanalyses with (guqh) and without (gvur) assimilation of altimeter-derived sea level anomalies. See text for details on the experiments

6.6 Multi-model ensembles

The benefit of multi-model combination for seasonal forecasts has been demonstrated since the early 2000s in Demeter (Palmer et al, 2006) and Ensemble projects (Weisheimer et al, 2009). Multi-model seasonal forecasts produce overall more skilful and reliable forecasts than the individual models. This led to an operational European multi-model (EUROSIP) and, more recently, operational multi-model seasonal forecasts from the Copernicus Climate Change Service (C3S). For medium-range forecasting, Hagedorn et al (2012) showed that a multi-model ensemble containing nine ensemble prediction systems (EPS) from the TIGGE archive did not improve on the performance of the best single-model, the ECMWF EPS. However, a reduced multi-model system, consisting of only the four best ensemble systems, provided by Canada, the USA, the United Kingdom and ECMWF, showed an improved performance. However, the ECMWF EPS was the main contributor for the improved performance of the multi-model ensemble; that is, if the multi-model system did not include the ECMWF contribution, it was not able to improve on the performance of the ECMWF EPS alone. These studies showed that the multi-model combination works best when the component models have comparable level of skill and display different model errors. For extended-range forecasts, the benefit of multi-model combination has not been assessed yet. The S2S re-forecast database is too inhomogeneous (most models have different start dates and frequency) to make a multi-model assessment possible. However, since June 2017, all the S2S data providers produce real-time forecasts every Thursday (it has been an important achievement of the S2S Project to homogenize the start dates of the real-time forecasts), making multi-models studies possible.

The impact of combining the S2S models has been assessed using all the S2S database real-time forecasts from 6 June 2017 to 1 November 2018, once a week. This represents a total of 74 cases. The real-time forecasts of each model have been calibrated using the re-forecasts of the same model to

produce weekly mean anomalies. The skill of multi-model combinations using the 11 S2S models has been compared to the skill of the individual models, with a special focus on the comparison with the ECMWF model. Results suggest that for variables or indices where the ECMWF model clearly outperforms all the other models, the multi-model combination does not outperform the ECMWF extended-range forecasts. This is the case for the MJO (Figure 42), for which none of the various multi-model combinations of the S2S models tested displayed higher skill scores than the ECMWF model. This is also the case for 500 hPa geopotential height skill scores (not shown). However, for some variables such as precipitation, multi-model combinations can produce more skilful and reliable extended-range forecasts (see for example Vigaud et al. (2017) and Figure 43). These results would support the potential benefit for users of an operational extended-range multi-model, similar to C3S or EUROSIP for long-range forecasting.

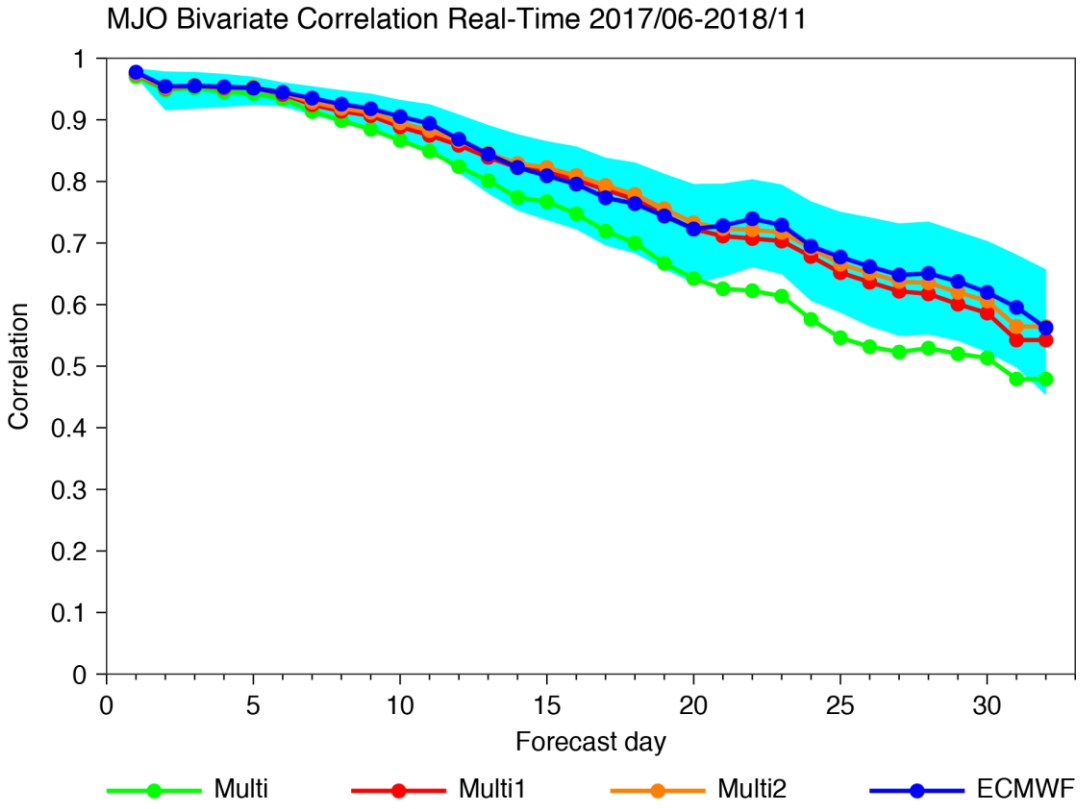

Figure 42: Evolution of the MJO bivariate correlation between the model ensemble means and ERA Interim as a function of lead time for the ECMWF real-time extended-range forecasts (dark blue line) produced every Thursday between 8 June 2017 and 1 November 2018 (74 cases). The cyan shaded area represents the 95% level of confidence computed from a 10,000 bootstrap re-sampling procedure. The other curves show the evolution of the MJO bivariate correlation of the equal-weight multi-model combination of 10 S2S models (green curve), weighted multi-model combination (red curve) with a weight proportional to the re-forecast skill, and the weighted combination of the 5 best models: ECMWF, UKMO, BoM, NCEP and CNRM (orange curve).

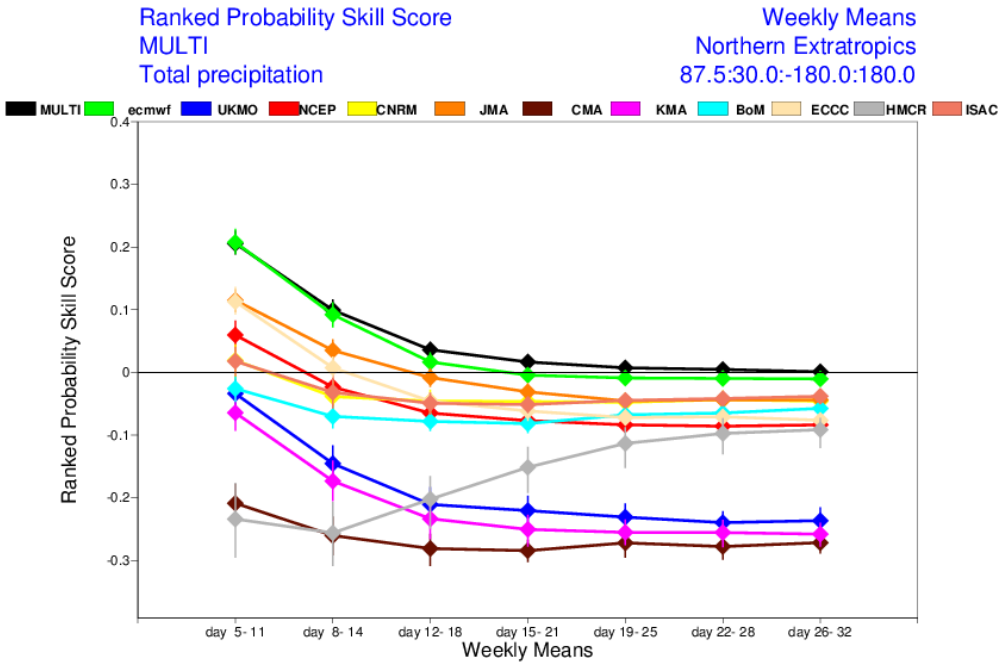

Figure 43: Ranked probability Skill Score (RPSS) of the weekly mean precipitation over the Northern Extratropics over the period 8 June 2017 to 1 November 2018 as a function of lead time for each S2S model. The black curve represents the scores of the equal-weight multi-model combination.

7 Conclusion

7.1 Summary

Extended-range forecasts have been produced operationally since 2004. The last SAC report on extended-range forecasting was issued in 2014. Since 2014, several important changes in the forecasting system have been introduced: new model cycles, increased horizontal resolution in the atmosphere and increased horizontal and vertical resolution in the ocean, interactive sea-ice, re-forecast initialization using ERA5. The extended-range forecast skill has improved since 2014, but mostly in the Tropics, with a noticeable gain of 3 days in the prediction of the MJO. In the Extratropics, there has been a slight, although not statistically significant, improvement in the NAO skill scores, but a slight degradation in the prediction of the stratospheric sub-seasonal variability. The use of ERA5 to initialize the re-forecasts led to an important improvement in the re-forecast probabilistic skill scores up to week 3 in the Extratropics and week 4 in the Tropics. This large improvement demonstrates the importance of the quality in atmospheric initial conditions for extended-range forecasts. The use of ERA5 also leads to more consistency between the real-time forecasts and re-forecasts, although the impact on the calibrated real-time skill scores is limited.

The establishment of the WWRP/WCRP S2S database has been an opportunity to compare for the first time the performance of the ECMWF extended-range forecasts with other centres. Results of the S2S model inter-comparison suggest that ECMWF has a strong lead for this time range. For instance, ECMWF MJO forecasts are more than a week more skilful than the second-best model. Other areas where ECMWF has an important lead include Euro-Atlantic weather regimes and sea-ice prediction. However, there are some areas where other S2S models outperform ECMWF such as the prediction of SSWs.

Despite the improvements since 2014, persistent systematic model errors and issues relating to the system configuration need to be addressed in order to maintain or increase the lead for the prediction at this time range:

- Difficulty of the MJO to propagate across the Maritime Continent
- Too weak MJO teleconnections over the Euro-Atlantic sector
- Biases in the decay of the Rossby Wave packets
- Large SST biases in the western boundary current regions.
- Inconsistencies between real-time and re-forecast initialization

We have tested a range of modelling options to understand and address these issues. The results suggest that improving on the delicate convection-wind-SST feedback in the Western Pacific equatorial region improves the propagation of the MJO. Furthermore, increasing the vertical resolution of the atmospheric model leads to larger skill score improvements, particularly in the Tropics, than increasing the horizontal resolution. Another potential source of skill improvement would be to change the current ensemble generation strategy from burst sampling to a lagged ensemble. Increasing the complexity of the earth system by including atmospheric chemistry such as interactive aerosols could also lead to improved extended-range forecasts as well as providing the possibility to issue new products. Further improvements could be achieved through a better initialization of the ocean and sea-ice. Section 6.5 showed that the use of sea-ice thickness observations can improve the prediction of sea-ice cover, and that coupled data assimilation could benefit extended-range forecasting. In the next 4 years, it is planned to test the impact of a 1/12 degree version of NEMO. It is hoped that this high-resolution ocean will produce more realistic boundary currents which should help reduce biases in these regions which are currently affecting the MJO teleconnections.

7.2 Proposed future system configurations and research

An important question for the coming year will be to decide the extended-range forecast configuration for the next HPC in Bologna. Since the characteristics of this new HPC were not known at the time of writing, no precise plan could be drafted at this stage. However, experiments with possible new configurations led to the following order of recommendations:

1. Run the real-time extended-range forecasts daily, instead of twice weekly, ideally with 51 ensemble members.
2. Run a small ensemble of re-forecasts with a daily or bi-daily frequency instead of twice weekly. Run a larger ensemble once a month for a more accurate skill assessment. This change could be done without increasing the computing cost.
3. Increase the vertical resolution to the same resolution as High Res, while benefitting from a 30-40% model speed up through the use of single precision.
4. Run the extended-range forecasts at low resolution from day 0.
5. Increase the horizontal resolution.

The changes 1 to 4 would not make it possible to increase the horizontal resolution as much as in the past (a factor of 4). A more modest horizontal increase (for example from Tco319 to Tco399) should make it possible to accommodate most of the changes proposed above. Although increasing horizontal resolution has a low impact on extended-range skill scores, it would still be beneficial in order to not increase too much the gap between the various operational systems.

In order to improve the consistency between real-time and re-forecasts, it would be desirable to:

- Explore the possibility of initializing the extended-range forecasts from a 4DVAR analysis produced at the same resolution as the first day of the forecast.
- Have a tighter link between the production of reanalysis and reforecast.
- Extend the operational re-forecasts to include the current year.

Research in the next 4 years will focus on the following themes, in addition to the research collaborations within the S2S project:

- Continue research on systematic errors.
- Continue research on tropical/extratropical, high-latitudes/mid-latitude and stratosphere/troposphere interactions.
- Continue research on addressing the western boundary current problems.
- Assess the impact of ocean observations on extended-range forecasts and make recommendations.
- Assess the impact of coupled data assimilation on extended-range forecasts.
- Continue aerosol/ozone experimentations.

A coordinated collaboration program that enables international research being transferred to operations will also be high in our agenda.

References

Ahlgrimm, M. and R. Forbes, 2014: Improving the representation of low clouds and drizzle in the ECMWF model based on ARM observations from the Azores. Mon. Wea. Rev., 142, 668-685. doi: 10.1175/MWR-D-13-00153.1

Balmaseda, M.A., K. Mogensen and A.T. Weaver, 2013: Evaluation of the ECMWF reanalysis system ORAS4. Q.J.R. Meterorol. Soc., 139: 1132–1161. doi: 10.1002/qj.2063

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

Balan-Sarojini, B., S. Tietsche, M. Mayer, M. Alonso-Balmaseda and H. Zuo, 2019: Towards Improved Sea Ice Initialization and Forecasting with the IFS, ECMWF Technical Memorandum 844. DOI: 10.21957/mt6m6rpwt

Baldwin, M.P. and T J Dunkerton, 2001: Stratospheric Harbingers of Anomalous Weather Regimes. Science, 294:581–584, 2001.

Bozzo, A., S. Remy, A. Benedetti, J. Flemming, P. Bechtold, M. Rodwell and J.-J. Morcrette, 2017: Implementation of a CAMS-based aerosol climatology in the IFS. ECMWF Technical Memorandum 801, 35 pp., https://www.ecmwf.int/sites/default/files/elibrary/2017/17219-implementation-cams-based-aerosol-climatology-ifs.pdf.

Boucher, O. and Coauthors, 2013: Clouds and aerosols. Climate Change 2013: The Physical Science Basis, T. F. Stocker et al., Eds., Cambridge University Press, 571–657.

Bougeault, P., Z. Toth, C. Bishop, B. Brown, D. Burridge, D.H. Chen, B. Ebert, M. Fuentes, T.M. Hamill, K. Mylne, J. Nicolau, T. Paccagnella, Y. Park, D. Parsons, B. Raoult, D. Schuster, P.S. Dias, R. Swinbank, Y. Takeuchi, W. Tennant, L. Wilson and S. Worley, 2010: The THORPEX Interactive Grand Global Ensemble. Bull. Amer. Meteor. Soc., 91, 1059–1072, https://doi.org/10.1175/2010BAMS2853.1

Buizza, R. and co-authors, 2018: The development and evaluation process followed at ECMWF to upgrade the Integrated Forecasting System (IFS). ECMWF Technical Memorandum, TM829, 47 pages.

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

DelSole, T. and M.K. Tippett, 2016: Forecast Comparison Based on Random Walks. Monthly Weather Review, 144(2), 615–626. https://doi.org/10.1175/MWR-D-15-0218.1

DeMott, C.A., J.J. Benedict, N.P. Klingaman, S.J. Woolnough and D.A. Randall, 2016: Diagnosing ocean feedbacks to the MJO: SST-modulated surface fluxes and the moist static energy budget, J. Geophys. Res. Atmos., 121, 8350– 8373, doi:10.1002/2016JD025098.

Ferranti L, L. Magnusson F. Vitart and D.S. Richardson, 2018: How far in advance can we predict changes in large-scale flow leading to severe cold conditions over Europe? Q.J.R. Meteorol Soc, 1–15. https://doi.org/10.1002/qj.3341

Ferro, C.A., 2014: Fair scores for ensemble forecasts. Q.J.R. Meteorol. Soc., 140: 1917-1923. doi:10.1002/qj.2270

Flemming, J. and Coauthors, 2017: The CAMS interim reanalysis of carbon monoxide, ozone and aerosol for 2003–2015. Atmos. Chem. Phys., 17, 1945–1983, https://doi.org/10.5194/acp-17-1945-2017. Crossref, Google Scholar

Forbes, R., 2018: Improved precipitation forecasts in IFS Cycle 45r1. ECMWF Newsletter No. 156, p4.

Goessling, H. F. and T. Jung, 2018: A probabilistic verification score for contours: Methodology and application to Arctic ice edge forecasts. Q.J.R. Meteorol. Soc., 144(712), 735–743. https://doi.org/10.1002/qj.3242

Goessling, H.F., S. Tietsche, J.J. Day, E. Hawkins and T. Jung, 2016: Predictability of the Arctic sea ice edge. Geophysical Research Letters, 43(4), 1642–1650. https://doi.org/10.1002/2015GL067232

Gottschalck, J. and Coauthors, 2010: A framework for assessing operational Madden–Julian Oscillation forecasts: A CLIVAR MJO Working Group project. Bull. Amer. Meteor. Soc., 91, 1247–1258, doi: https://doi.org/10.1175/2010BAMS2816.1.

Grazzini, F. and F. Vitart, 2015: Atmospheric predictability and Rossby wave packets. Q.J.R. Meteorol. Soc, 141, 2793–2802.

Guo, Y., B. Tian, R.A. Kahn, O. Kalashnikova, S. Wong and D.E. Waliser, 2013: Tropical Atlantic dust and smoke aerosol variations related to the Madden–Julian oscillation in MODIS and MISR observations. J. Geophys. Res. Atmos., 118, 4947–4963, https://doi.org/10.1002/jgrd.50409.

Hagedorn, R. Buizza, R. Hamill, T.M. Leutbecher and Palmer, T.N., 2012: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. Q.J.R. Meteorol. Soc., 138: 1814-1827. doi:10.1002/qj.1895

Hogan, R.J. and A. Bozzo, 2018: A flexible and efficient radiation scheme for the ECMWF model. Journal of Advances in Modeling Earth Systems, 10, 1990–2008. https://doi.org/10.1029/2018MS001364

Jie, W., F. Vitart, T. Wu and X. Liu, 2017: Simulations of the Asian summer monsoon in the sub-seasonal to seasonal prediction project (S2S) database. Q.J.R. Meteorol. Soc., 143: 2282-2295. doi:10.1002/qj.3085

Keeley SPE, R.T. Sutton, L.C. Shaffrey, 2012: The impact of North Atlantic sea surface temperature errors on the simulation of North Atlantic European region climate. Q.J.R. Meteorol. Soc., 138(668):1774–1783. https://doi.org/10.1002/qj.1912

Kim, H., D. Kim, F. Vitart, V.E. Toma, J. Kug and P.J. Webster, 2016: MJO Propagation across the Maritime Continent in the ECMWF Ensemble Prediction System. J. Climate, 29, 3973–3988, https://doi.org/10.1175/JCLI-D-15-0862.1

Koster, R.D., S.P.P. Mahanama, T.J. Yamada, Gianpaolo Balsamo, A.A. Berg, M. Boisserie,P.A. Dirmeyer, F.J. Doblas-Reyes, G. Drewitt, C.T. Gordon, Z. Guo, J.-H. Jeong, W.-S. Lee, Z. Li, L. Luo, S. Malyshev, W.J. Merryfield, S.I. Seneviratne, T. Stanelle, B.J.J.M. van den Hurk, F. Vitart, and E.F. Wood, 2011: The Second Phase of the Global Land–Atmosphere Coupling Experiment: Soil Moisture Contributions to Subseasonal Forecast Skill. J. Hydrometeor, 12, 805–822.

Laloyaux, P., M. Balmaseda, D. Dee, K. Mogensen and P. Janssen, 2016: A coupled data assimilation system for climate reanalysis. Q.J.R. Meteorol. Soc., 142, 65–78.

Lee, R.W., T.J. Woollings, B.J. Hoskins, K.D. Williams, C.H. O'Reilly and G. Masato, 2018: Impact of Gulf Stream SST biases on the global atmospheric circulation. Climate Dynamics. ISSN 1432-0894 doi: https://doi.org/10.1007/s00382-018-4083-9

Lee, C.-Y., Suzana J. Camargo, Fréderic Vitart, Adam H. Sobel and Michael K. Tippett, 2018: Subseasonal Tropical Cyclone Genesis Prediction and MJO in the S2S Dataset. Weather and Forecasting 33:4, 967-988.

Lin, H. and G. Brunet, 2018: Extratropical Response to the MJO: Nonlinearity and Sensitivity to the Initial State. J. Atmos. Sci., 75, 219–234, https://doi.org/10.1175/JAS-D-17-0189.1

Liu, C., B. Tian, K.F. Li, G.L. Manney, N.J. Livesey, Y.L. Yung, and D.E. Waliser, 2014: Northern Hemisphere mid-winter vortex-displacement and vortex-split stratospheric sudden warmings: Influence of the Madden-Julian Oscillation and Quasi-Biennial Oscillation. Journal of Geophysical Research: Atmospheres, 119, 12,599– 12,620. https://doi.org/10.1002/2014JD021876

Lock, S-J, Lang, STK, Leutbecher, M, Hogan, RJ, Vitart, F. Treatment of model uncertainty from radiation by the Stochastically Perturbed Parametrization Tendencies (SPPT) scheme and associated revisions in the ECMWF ensembles. Q J R Meteorol Soc. 2019; 145 (Suppl. 1): 75- 89. https://doi.org/10.1002/qj.3570

Miao, Q., B. Pan, H. Wang, K. Hsu, S. Sorooshian, 2019: Improving Monsoon Precipitation Prediction Using Combined Convolutional and Long Short-Term Memory Neural Network. Water 2019, 11, 977.

Morcrette, J.-J. and Coauthors, 2009: Aerosol analysis and forecast in the European Centre for Medium-Range Weather Forecasts Integrated Forecast System: Forward modelling. J. Geophys. Res., 114, D06206, https://doi.org/10.1029/2008JD011235.

Nakamura H, T. Sampe, A. Goto, W. Ohfuchi, S.P. Xie, 2008: On the importance of midlatitude oceanic frontal zones for the mean state and dominant variability in the tropospheric circulation. Geophys. Res Lett 35(15): L15,709. https://doi.org/10.1029/2008GL034010

Palmer, T.N., A. Alessandri, U. Andersen, P. Cantelaube, M. Davey, P. Délécluse, M. Déqué, E. Díez, F.J. Doblas-Reyes, H. Feddersen, R. Graham, S. Gualdi, J.-F. Guérémy, R. Hagedorn, M. Hoshen, N. Keenlyside, M. Latif, A. Lazar, E. Maisonnave, V. Marletto, A.P. Morse, B. Orfila, P. Rogel, J.-M. Terres, M.C. Thomson, 2004. Development of a European multi-model ensemble system for seasonal to inter-annual prediction (DEMETER). Bulletin of the American Meteorological Society, 85, 853-872.

Polichtchouk, I., T. Stockdale, P. Bechtold, M. Diamantakis, S. Malardel, I. Sandu, F. Vana and N. Wedi, 2019: Control on stratospheric temperature in IFS: resolution and vertical advection. ECMWF Technical Memorandum, 427, 36pp.

Quinting, J.F. and F. Vitart, 2019: Representation of synoptic-scale Rossby wave packets and blocking in the S2S prediction project database. Geophysical Research Letters, 46, 1070– 1078. https://doi.org/10.1029/2018GL081381

Rashid, H.A., H.H. Hendon, M.C. Wheeler and O. Alves, 2011: Prediction of the Madden–Julian oscillation with the POAMA dynamical prediction system. Climate Dyn., 36, 649–661, doi:https: //doi.org/10.1007/s00382-010-0754-x.

Reddy, M.S., O. Boucher, N. Bellouin, M. Schulz, Y. Balkanski, J.-L. Dufresne and M. Pham, 2005: Estimates of global multicomponent aerosol optical depth and direct radiative perturbation in the Laboratoire de Météorologie Dynamique general circulation model. J. Geophys. Res., 110, D10S16, https://doi.org/10.1029/2004JD004757.

Roberts, C., F. Vitart. M.A. Balmaseda and F. Molteni, 2019: The atmospheric response to increased ocean model resolution in the ECMWF Integrated Forecasting System: a seamless approach. Submitted to J. Clim.

Rodwell, M.J. and T. Jung, 2008: Understanding the local and global impacts of model physics changes: An aerosol example. Q.J.R. Meteorol. Soc., 134, 1479–1497, https://doi.org/10.1002/qj.298.

Röthlisberger, M., O. Martius and H. Wernli, 2018: Northern Hemisphere Rossby wave initiation events on the extratropical jet—A climatological analysis. Journal of Climate, 31(2), 743–760. https://doi.org/10.1175/JCLI-D-17-0346.1

Scaife A.A., D. Copsey, C. Gordon, C. Harris, T. Hinton, S. Keeley, A. O'Neill, M. Roberts, K. Williams, 2011: Improved Atlantic winter blocking in a climate model. Geophys Res Lett. https://doi.org/10.1029/2011GL049573

Scoccimarro, E., A Bellucci, A Storto, S Gualdi, S Masina and A Navarra, 2018: Remote subsurface ocean temperature as a predictor of Atlantic hurricane activity. Proceedings of the National Academy of Sciences 115 (45), 11460-11464

Schepers, D., E. de Boisséson, R. Eresmaa, C. Lupu and P. de Rosnay, 2018: CERA-SAT: A coupled satellite-era reanalysis. ECMWF Newsletter 155.

Seo, K. and H. Lee, 2017: Mechanisms for a PNA-Like Teleconnection Pattern in Response to the MJO. J. Atmos. Sci., 74, 1767–1781, https://doi.org/10.1175/JAS-D-16-0343.1

Son, S.-W., Y. Lim, C. Yoo, H.H. Hendon and J. Kim, 2017: Stratospheric control of the Madden–Julian oscillation. Journal of Climate, 30(6), 1909– 1922. https://doi.org/10.1175/JCLI-D-16-0620.1

Taguchi, M., 2018: Comparison of sub-seasonal-to-seasonal model forecasts for major stratospheric sudden warmings. Journal of Geophysical Research: Atmospheres, 123. https://doi.org/10.1029/2018JD028755

Tegen, I., P. Hollrig, M. Chin, I. Fung, D. Jacob and J. Penner, 1997: Contribution of different aerosol species to the global aerosol extinction optical thickness: Estimates from model results. J. Geophys. Res., 102, 23 895–23 915, https://doi.org/10.1029/97JD01864.

Tian, B. and Coauthors, 2008: Does the Madden-Julian oscillation influence aerosol variability? J. Geophys. Res., 113, D12215, https://doi.org/10.1029/2007JD009372.

Tian, B., D.E. Waliser, R.A. Kahn and S. Wong, 2011: Modulation of Atlantic aerosols by the Madden-Julian oscillation. J. Geophys. Res., 116, D15108, https://doi.org/10.1029/2010JD015201.

Tompkins, A. M., C. Cardinali, J.-J. Morcrette and M. Rodwell, 2005: Influence of aerosol climatology on forecasts of the African easterly jet. Geophys. Res. Lett., 32, L10801, https://doi.org/10.1029/2004GL022189.

Tripathi, O., A. Charlton-Perez, M. Sigmond and F. Vitart, 2015: Enhanced long-range forecast skill in boreal winter following stratospheric strong vortex conditions. Environmental Research Letters 10/2015; 10. DOI:10.1088/1748-9326/10/10/104007

Tseng, K., E. Maloney and E. Barnes, 2019: The Consistency of MJO Teleconnection Patterns: An Explanation Using Linear Rossby Wave Theory. J. Climate, 32, 531–548, https://doi.org/10.1175/JCLI-D-18-0211.1

Vidard, A., M. Balmaseda and D. Anderson, 2008: Assimilation of altimeter data the ECMWF ocean analysis system. Mon. Wea. Rev, 137, 1393-1408.

Vigaud, N., A.W. Robertson and M.K. Tippett, 2017: Multi-model Ensembling of Subseasonal Precipitation Forecasts over North America. Mon. Wea. Rev., 145, 3913–3928, https://doi.org/10.1175/MWR-D-17-0092.1

Vitart, F., 2014: Evolution of ECMWF sub-seasonal forecast skill scores. Q.J.R. Meteorol. Soc., 140: 1889-1899. doi:10.1002/qj.2256

Vitart, F. 2017: Madden-Julian Oscillation Prediction and Teleconnections in the S2S Database. Q.J.R. Meteorol. Soc., 143: 2210–2220. doi:10.1002/qj.3079

Vitart, F., G. Balsamo, J.R. Bidlot, S. Lang, I. Tsonevsky, D. Richardson, M. Alonso-Balmaseda, 2019: Use of ERA5 to Initialize Ensemble Re-forecasts. ECMWF Technical Memorandum, 841, 14pp.

Vitart, F., C. Ardilouze, A. Bonet, A. Brookshaw, M. Chen, C. Codorean, M. Déqué, L. Ferranti, E. Fucile, M. Fuentes, H. Hendon, J. Hodgson, H. Kang, A. Kumar, H. Lin, G. Liu, X. Liu, P. Malguzzi, I. Mallas, M. Manoussakis, D. Mastrangelo, C. MacLachlan, P. McLean, A. Minami, R. Mladek, T. Nakazawa, S. Najm, Y. Nie, M. Rixen, A.W. Robertson, P. Ruti, C. Sun, Y. Takaya, M. Tolstykh, F. Venuti, D. Waliser, S. Woolnough, T. Wu, D. Won, H. Xiao, R. Zaripov and L. Zhang, 2017: The Subseasonal to Seasonal (S2S) Prediction Project Database. Bull. Amer. Meteor. Soc., 98, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1

Vitart, F. and M. Alonso-Balmaseda, 2018: Impact of sea surface temperature biases on extended-range forecasts, ECMWF Technical Memorandum, 830, 19pp.

Weigel, A.P., M.A. Liniger and C. Appenzeller, 2007: Generalization of the Discrete Brier and Ranked Probability Skill Scores for Weighted Multimodel Ensemble Forecasts. Mon. Wea. Rev., 135, 2778–2785, https://doi.org/10.1175/MWR3428.1

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

Woollings, T., B Hoskins, M Blackburn, D Hassell and K Hodges, 2010: Storm track sensitivity to sea surface temperature resolution in a regional atmosphere model. Climate dynamics 35 (2-3), 341-353

Woolnough, S. J., F. Vitart and M. A, Balmaseda, 2007: The role of the ocean in the Madden-Julian Oscillation: Implications for MJO prediction Q.J.R. Meteorol. Soc., 133, 117-128.

Yadav, P. and D.M. Straus, 2017: Circulation Response to Fast and Slow MJO Episodes. Mon. Wea. Rev., 145, 1577–1596, https://doi.org/10.1175/MWR-D-16-0352.1

Zampieri, L., H.F. Goessling and T. Jung, 2018: Bright prospects for Arctic sea ice prediction on subseasonal time scales. Geophysical Research Letters, 45, 9731– 9738. https://doi.org/10.1029/2018GL079394

Zuo, H., M.A. Balmaseda and K. Mogensen, 2015: The new eddy-permitting ORAP5 ocean reanalysis: description, evaluation and uncertainties in climate signals. Clim. Dyn. 10.1007/s00382-015-2675-1

Zuo, H., M.A. Balmaseda, S. Tietsche, K. Mogensen and M. Mayer, 2019: The ECMWF operational ensemble reanalysis-analysis system for ocean and sea-ice: a description of the system and assessment. Ocean Sci., 15, 779–808. https://doi.org/10.5194/os-15-779-2019