H2020 - Research and Innovation Action

APPLICATE

Advanced Prediction in Polar regions and beyond: Modelling, observing system design and LInkages associated with a Changing Arctic climaTE

Grant Agreement No: 727862

Deliverable No. 5.2

Strengths and limitations of state-of-the-art weather and climate prediction systems

Submission of Deliverable

Work Package	WP5 Improved predictive capacity
Deliverable No	**5.2**
Deliverable title	**Strengths and limitations of state-of-the-art weather and climate prediction systems**
Version	**2**
Status	Final
Dissemination level	PU - Public
Lead Beneficiary	11 - CNRS-G
Contributors	**□ 1 – AWI ■ 2 – BSC ■ 3 - ECMWF** **□ 4 – UiB □ 5 – UNI Research ■ 6 – MET Norway** **■ 7 – Met Office ■ 8 – UCL □ 9 - UREAD** **□ 10 – SU ■ 11 – CNRS-GAME □ 12 - CERFACS** **□ 13 – AP □ 14 – UiT □ 15 - IORAS** **□ 16 - MGO**
Due Date	31/10/18
Delivery Date	24/10/18
Coordinating author	**Lauriane Batté (lauriane.batte@meteo.fr) (CNRS)**
Contributing authors	**Juan Camilo Acosta Navarro (BSC)** **Morten Koltzow (MET Norway)** **Linus Magnusson (ECMWF)** **Pablo Ortega (BSC)** **Leandro Ponsoni (UCL)** **Doug Smith (Met Office)**

This project has received funding from the European Union's Horizon 2020 Research & Innovation programme under grant agreement No. 727862.

Table of Contents

EXECUTIVE SUMMARY

This document provides an overview of predictive capacity over the Arctic and mid-latitudes of current state-of-the-art prediction systems ranging from numerical weather prediction (NWP) to seasonal time scales.

The assessment is mainly based on forecasting systems and climate models contributing to the APPLICATE project. This deliverable therefore provides a thorough evaluation of the forecast models included in the WP5 stream 1 experiments, and a baseline for future improvements to current systems resulting from developments in the framework of the project.

Beyond commonly used verification metrics for the evaluation of weather and climate predictions, illustrations of current systems predictive capacity are shown by focusing on specific phenomena and case studies (e.g. extreme rainfall on Svalbard). With the perspective of providing useful and reliable forecasts for potential end-users, some skill evaluations on more user-relevant metrics were included.

Results on the weather prediction time scales show the impact of horizontal resolution in better representing precipitation extremes, although some weaknesses remain in a 2.5 km resolution configuration for the Svalbard case study examined in this deliverable. More generally, high resolution limited area models show added value with respect to global models depending on the parameter and region of interest.

At the medium range (5 days), the evaluation of the European Centre for Medium-range Weather Forecasts (ECMWF) forecasts over 1990-present for geopotential height at 500 hPa shows that these have been steadily improving over the Arctic, at the same rate as the Northern Hemisphere in general. Skill and biases are found to vary according to the region and season of interest.

Seasonal re-forecasts over a common 1993-2014 period were evaluated for both atmospheric and sea ice concentration fields. The skill of the systems is quite limited, consistent with previous works. For sea ice, forecast performance for boreal summer seems to depend quite strongly on systematic errors which appear in some systems from the initialization time step. This deliverable also presents results from a statistical forecasting framework, using HighResMIP model simulations to evaluate lagged predictability of sea ice volume with sea ice volume and area as predictors. It appears from the results presented that sea ice area does not add much additional predictability to the information provided by sea ice volume.

1. INTRODUCTION

1.1 Background and objectives

This deliverable provides a comprehensive overview of current prediction capabilities in state-of-the-art weather and climate forecasting models. The report synthesizes findings from the analysis of operational systems as well as the first stream of experiments performed in the first 18 months of the APPLICATE project (stream 1, see deliverable 5.1).

The content of this deliverable results from work led in task 5.2, which aims to assess several aspects of the quality of forecasting systems existing at the start of the APPLICATE project. Results presented in this deliverable will serve as a baseline for testing improved performance of the systems developed during the second phase of the project, building on work led in WP2 as well as in task 5.3.

1.2 Organisation of this report

Rather than presenting an exhaustive list of scores for numerous atmospheric and sea ice variables, this report focuses on some key aspects of forecast quality, working from the local and numerical weather prediction scales to the regional and global scale for the longer ranges.

The report is organised as follows: part 2 provides detailed information on the methodology used, including the models, reference data, and skill metrics chosen. Part 3 presents a detailed analysis of forecast quality at different time ranges for representing extreme rainfall events over Svalbard. We then take a step back and present indicators of model performance at the regional level (Arctic and/or Northern Hemisphere midlatitudes), for both the medium range (part 4) and the seasonal time scales (part 5). Part 5 also incorporates results on statistical prediction of sea ice volume based on HighResMIP simulations. We summarize key points in a conclusions section (part 6).

2. METHODOLOGY

2.1. Model and reference data

NWP and medium-range forecasts

AROME Arctic (AA) is a limited area NWP model in operational use for northern Norway, the Svalbard region and the Barents Sea (blue frame in Figure 4.2.1) operated by the Norwegian Meteorological Institute (MET Norway). The model is based on the High Resolution Limited Area Model (HIRLAM)–ALADIN Research on Mesoscale Operational NWP in Europe (HARMONIE) AROME configuration (Bengtsson et al. 2017). AA use a 3D-var assimilation scheme for upper air and optimal interpolation for surface analysis. The model has 2.5km horizontal grid spacing and 65 vertical levels. As AA is not a global model it uses ECMWF high-resolution forecasts (see below) as lateral boundary conditions (more details are found in Müller et al., 2017). Analysis of several years of AA forecasts contributes to “Multi-Scale predictions of extremes: Rainfall Svalbard” (section 3.1, 3.3) and “Evaluation of short- and medium-range forecast over the Arctic (section 4.2).

ECMWF produces global forecasts aimed at medium-range through to sub-seasonal and seasonal scales. The deterministic high-resolution forecast (HRES hereafter) uses 9 km horizontal resolution and 137 vertical levels (of which 20 are below 1000 metres) and runs twice a day out to a lead time of 10 days. The ensemble (ENS) uses 18 km horizontal resolution and 91 vertical levels and runs out to 15 days twice a day. Twice a week (Monday and Thursday) the ensemble forecasts are extended to 45 days at 36 km resolution to provide forecasts on the sub-seasonal time scale. Once a month, seasonal forecasts are produced with 7-month lead time using ECMWF SEAS5 system. The atmospheric component of the forecasting system is the ECMWF Integrated Forecasting System (IFS) model (https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model/ifs-documentation). Since June 2018, all systems are coupled to the Nemo (Nucleus for European Modelling of the Ocean) ocean model with 0.25 degree resolution (in June 2018 HRES became coupled, before only ENS and SEAS5 were coupled). The ocean coupling is described in Mogensen et al. (2017) and references therein.

The initial conditions are created separately for the atmosphere and ocean. The atmospheric initial conditions are produced with a 4-dimensional variational data assimilation (4D-Var, Rabier et. al, 2000). To provide background error statistics a 25-member ensemble of 4D-Var assimilations (EDA) is run with a lower (18 km) resolution (Bonavita et al., 2012). The EDA members are also used to initialise the ensemble forecast (see below). The ocean initial conditions are provided by the ECMWF OCEAN5 operational 3-dimensional variational data assimilation system (Zuo et al., 2018).

Ensemble forecasts are run with the aim of estimating range of possible future states (Leutbecher and Palmer, 2008). At ECMWF, the ensemble consists of 50 perturbed forecasts and 1 unperturbed member with the same resolution (control member). The ensemble is generated by applying initial perturbations based on a combination of EDA and singular vector perturbations and model uncertainties represented by the Stochastically Perturbed Parametrization Tendency scheme and the Stochastic Kinetic Energy Backscatter scheme (SPPT and SKEB, see Leutbecher et al.(2017) for details).

Seasonal re-forecasts

Table 2.1 presents information on the seasonal re-forecasts evaluated in this study (in section 5). The following paragraphs provide more detailed information on the different coupled models.

CNRM-CM6-1 is the global coupled model developed by CNRM and CERFACS for CMIP6 (Voldoire et al. 2018). The coupled model uses ARPEGE-Climate v6.3 for the atmosphere and NEMO3.6 - GELATOv6 for the ocean and sea ice. The land surface component is SURFEXv8. Coupling between atmosphere/land and ocean is called in the SURFEX interface using the OASIS-MCT code. This global coupled model (GCM) was used to run seasonal re-forecast experiments initialized in May and November 1993-2014 as part of APPLICATE stream 1 (see Deliverable 5.1).

EC-Earth3.2 is based on the ECMWF's atmospheric circulation model IFS, cycle 36r4 and the land surface model H-Tessel. The ocean component is a recent version of the ocean model NEMO3.6 and the sea-ice model is a recent version of the Louvain-la-Neuve Sea Ice Model (LIM3). The different components communicate via the coupler OASIS-3. EC-Earth3.2

is used for seasonal re-forecast experiments initialized in May and November 1993-2014 and for climate change experiments covering the period 1950-2014 (1950-2050 once the forcing data is released).

Seasonal re-forecasts from GloSea5 (MacLachlan et al. 2015) were also included in the analysis. GloSea5 is the UK Met Office operational seasonal forecast system, using the Met Office Unified Model (UM) global atmosphere model coupled to the NEMO ocean model with CICE sea ice at a 0.25° horizontal resolution.

So as to complete our evaluation with another operational state-of-the-art seasonal forecasting system, we also included the ECMWF SEAS5 re-forecasts (see full description above) in our analysis of sea ice seasonal forecasts. Both GloSea5 and SEAS5 are part of the Copernicus Climate Changes Services operational seasonal forecasting systems, while CNRM-CM is a different model version to that contributing to Copernicus.

Model/System	**CNRM-CM6-1**	**SEAS5**	**GloSea5**	**EC-Earth 3.2.2**
Atmosphere	ARPEGE 6.3	IFS Cy43r1	UM v6	IFS Cy36r4
Ocean	NEMO 3.6	NEMO 3.4	NEMO 3.4	NEMO 3.6
Sea ice	GELATO v6	LIM2	CICE 4.1	LIM3
Atmospheric resolution	tl127l91r (~ 1.4°)	TCo319L91	N216L85	tl255l91r (~ 0.7°)
Ocean resolution	eORCA1 L75	ORCA 0.25 L75	ORCA 0.25 L75	ORCA1L75
Initial conditions	GLORYS (Mercator)	ORS-S5	NEMOVAR	Forced NEMO run
Ensemble size	30	25	28*	10

*Tab. 2.1: Characteristics of the seasonal re-forecasts included in the analysis presented in section 5. All systems are initialized in the atmosphere with ERA-Interim. * All re-forecasts are initialized on the 1st of the month, except for GloSea5 for which 7 members from the 9th, 17th, 25th of the previous month as well as 7 from the 1st of the initialization month are grouped into a 28-member ensemble.*

Climate change simulations

This deliverable initially envisaged the analysis of climate change simulations with EC-Earth3 and ECHAM6-FESOM. These included the production of transient experiments for the period 1950-2050, and present-day control simulations with fixed forcing from year 1950. However, a major delay in the generation by the CMIP6 community of the radiative forcings for the future scenarios has prevented us from starting the transient simulations. Also, a potential bug over the Arctic has been identified in the control experiment with ECHAM6-FESOM, that is currently under investigation. Due to these problems, and given the specific focus of this deliverable on weather and climate prediction, we have decided to exclude the climate change simulations from the report. They will be covered extensively in the final Deliverable

5.6 on the integrated added-value of APPLICATE on weather and climate predictions and projections.

Reference datasets

High-resolution short-range forecasts over Svalbard are compared to station data over six locations of the region for daily precipitation. Evaluations of NWP predictions are also presented with respect to data from 120 SYNOP stations in Norway.

For the seasonal re-forecast evaluations, precipitation was compared to the GPCP v2.2 monthly precipitation analysis dataset (Adler et al. 2003). The ERA-Interim reanalysis (Dee et al. 2011) is used for most evaluations of other atmospheric variables.

Monthly mean sea ice concentration data based on brightness temperature (Cavalieri et al. 1996) was retrieved from the National Snow and Ice Data Center (NSIDC).

2.2. Metrics of forecast quality

2.2.1. Deterministic scores and metrics

Atmospheric and sea ice fields

Below we give a list of metrics applied in this report and a short description of them. Where no other references are given, they are explained in more detail at http://www.cawcr.gov.au/projects/verification/.

Mean Error (ME)
Also called bias. Measures the average or systematic error, and indicates the sign of the error. Negatively oriented score (best score 0).

Mean Absolute Error (MAE)
Measures the average of the magnitude of the error. Negatively oriented score (best score 0).

Root Mean Square Error (RMSE)
Measures the deviation from the observed values, but puts more weight on larger errors. Negatively oriented score (best score 0).

Correlation
Measures the linear association between forecasts and observations. Does not take bias into account. Positively oriented score (best score 1).

Equitable Threat Score (ETS)
Measures the fraction of observed and/or forecast events that were correctly predicted, adjusted for hits associated with random chance. Positively oriented score (best score 1).

Frequency bias
Measures the ratio of the frequency of forecasted events and the frequency of observed events. Best score 1, above (below) 1 indicates forecasting too many (few) events compared to the observations.

Ratio of Predictable Components (RPC)
Based on Eade et al. (2014), the RPC compares the predictable component in observations (PCobs) to that in model hindcasts (Pcmod). It will be used exclusively for the North Atlantic Oscillation Index (NAOI). It is in practice estimated as a lower bound using the right term in

the equation below, where r is the Pearson correlation coefficient divided by the square root of the ratio in the variance of ensemble mean NAOI and the mean of the variances of the NAOI for individual members.

$$RPC=\frac{PC_{obs}}{PC_{mod}}\geqslant\frac{r}{\sqrt{\sigma_{sig}^2/\sigma_{tot}^2}}$$

Sea ice edge

IIEE and decomposition

Integrated Ice Edge Error (IIEE, Goessling et al. 2016) is computed to evaluate the total spatial extent of errors in the position of the sea ice edge. The IIEE is the sum of areas where the presence of sea ice, defined with a 15% SIC threshold, is overestimated (O) and underestimated (U) with respect to reference data.

The IIEE can be decomposed into two terms, namely misplacement error (ME) and absolute extent error (AEE), as follows:

$$IIEE=O+U=|O-U|+2\cdot min(O,U)=AEE+ME$$

The absolute error corresponds to the total Pan-Arctic SIE error when this metric is computed over the region, whereas the misplacement error shows the compensation between areas with overestimation and areas with underestimation.

2.2.2. Probabilistic scores

Atmospheric and sea ice fields

Fair Continuous Ranked Probability Skill Score (FCRPSS)

The FCRPSS is the ensemble-size corrected, integrated squared difference between the cumulative distribution function of the forecasts and the corresponding value in the observations, compared with its climatological equivalent. I.e. the FCRPSS estimates the added value of a forecasting system over a climatological forecast. Values of one indicate a perfect forecast, while positive, zero and negative values indicate an improvement, no improvement and a degradation of the forecast over the climatology, respectively. The FCRPSS was estimated for each variable's winter average (DJF) at each grid point for each startdate of the re-forecast period. The FCRPSS map is estimated over 22 independent events and the climatology of the 22 years hindcast period is used as the reference forecast.

Sea ice edge

Spatial Probability Score (SPS)

A natural extension to the IIEE is used to examine skill of probabilistic forecasts for presence of sea ice at a grid point level. Goessling and Jung (2018) recently introduced the Spatial Probability Score (SPS) which consists in a spatial integral of the Brier Score for the probabilistic event of SIC exceeding the 15% threshold. With NSIDC data as a reference, the SPS is formulated as follows:

$$SPS=\iint\left(P_{SIC_f>0.15}(x,y)-1_{SIC_o>0.15}(x,y)\right)^2 dx\,dy$$

In this deliverable probabilities are computed by counting the fraction of ensemble members exceeding the 15% concentration threshold, and then bias-corrected using leave-one-out cross-validation.

3. MULTI-SCALE PREDICTIONS OF EXTREMES: RAINFALL IN SVALBARD

3.1. Introduction

On 7-10 November 2016 Svalbard was hit by extreme rainfall. For example Ny-Ålesund observed 86.8mm in 24hr (November normal 33mm) and 41.7mm in 24hr was observed at Svalbard Airport (November normal 15mm). Several landslides and (slush) avalanches were identified by satellite images at Svalbard during these days, even if it is believed that a frozen surface and modest surface snow amounts stabilized the situation. Within short time after the rainfall the temperature fell well below 0°C with the potential of producing ground ice and reducing availability of food for the wildlife. Evacuation of parts of Longyearbyen was done in advance (see for instance the following websites: https://titan.uio.no/node/2009, https://www.dagbladet.no/nyheter/130-evakuert-pa-svalbard-deler-av-longyearbyen-er-sperret-av-og-det-er-innfort-ferdselsforbud/64354331,https://norut.no/nb/news/kartlegging-med-radarsatellitt-gir-bedre-snoskredvarsling-og-beredskap, all in Norwegian). In general, such rain on snow events has a substantial impact on infrastructure, society and wildlife as described in more detail in Serreze et al. (2015) and Hansen et al. (2014). They are therefore one of the key types of extreme events to predict on different time-scales.

The extreme rainfall was caused by a warm and humid air stream from the south. Such streams are often referred to as atmospheric rivers (http://glossary.ametsoc.org/wiki/Atmospheric_river) and its signature can be seen in the vertically integrated water vapour transport (IVT) on the 8 November 2016 towards Svalbard (Figure 3.1.1). The figure is based on the ECMWF forecast product Extreme Forecast Index (EFI) for IVT, as described in Lavers et al. (2017). At the same time the temperature was extremely warm (Figure 3.1.1), and because of that the precipitation fell partly in form of rainfall. In this type of flow situations the precipitation is enhanced by orography. Such a process is inherently dependent on the model resolution to be captured in numerical weather forecasts. Model topography from ECMWF HRES (~ 9 km grid spacing) and AROME Arctic (2.5km grid spacing) is shown in Figure 3.1.2 together with the 24hr accumulated precipitation forecasts. The more detailed topography in AROME Arctic is a prerequisite to be able to forecast local variations in precipitation.

To predict extreme precipitation and heat in the European sector of the Arctic on different time scales involves different challenges. For short-range predictions, the model resolution needs to be sufficient to resolve the orography to capture the precipitation maximum. In medium-range predictions, with coarser resolution model systems, features like atmospheric rivers and large-scale patterns leading to strong advection of heat and moisture from lower latitudes need to be captured. In a statistical sense, such events are related to the Scandinavian blockings (Serreze et al., 2015). It is therefore important to capture the statistical properties of such patterns in (sub) seasonal forecasts and also in climate projections. If the blocking frequency is changing in a future climate, the statistics of extremes in the Arctic will do as well.

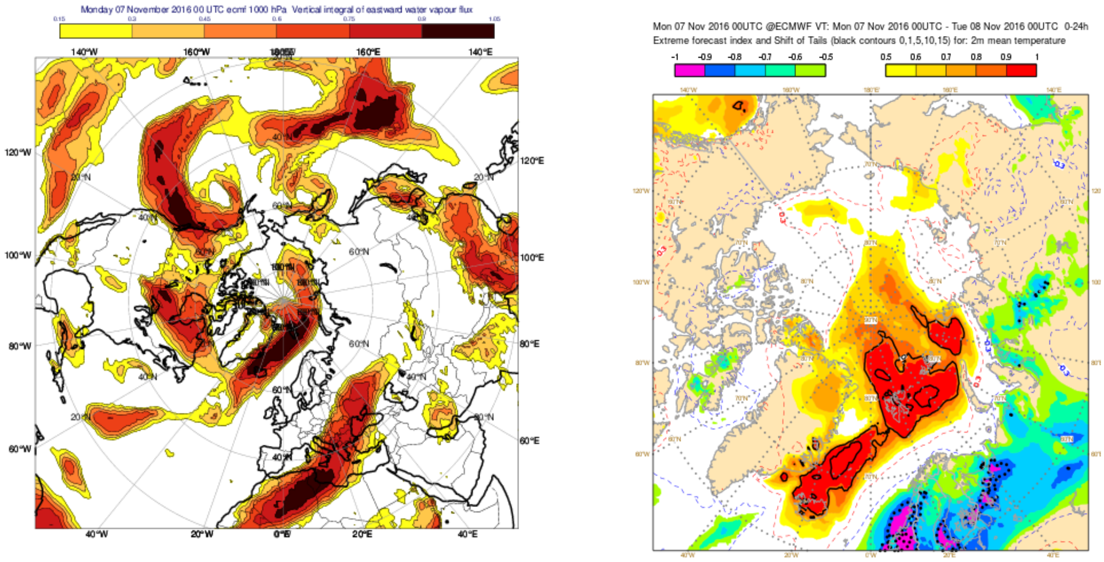
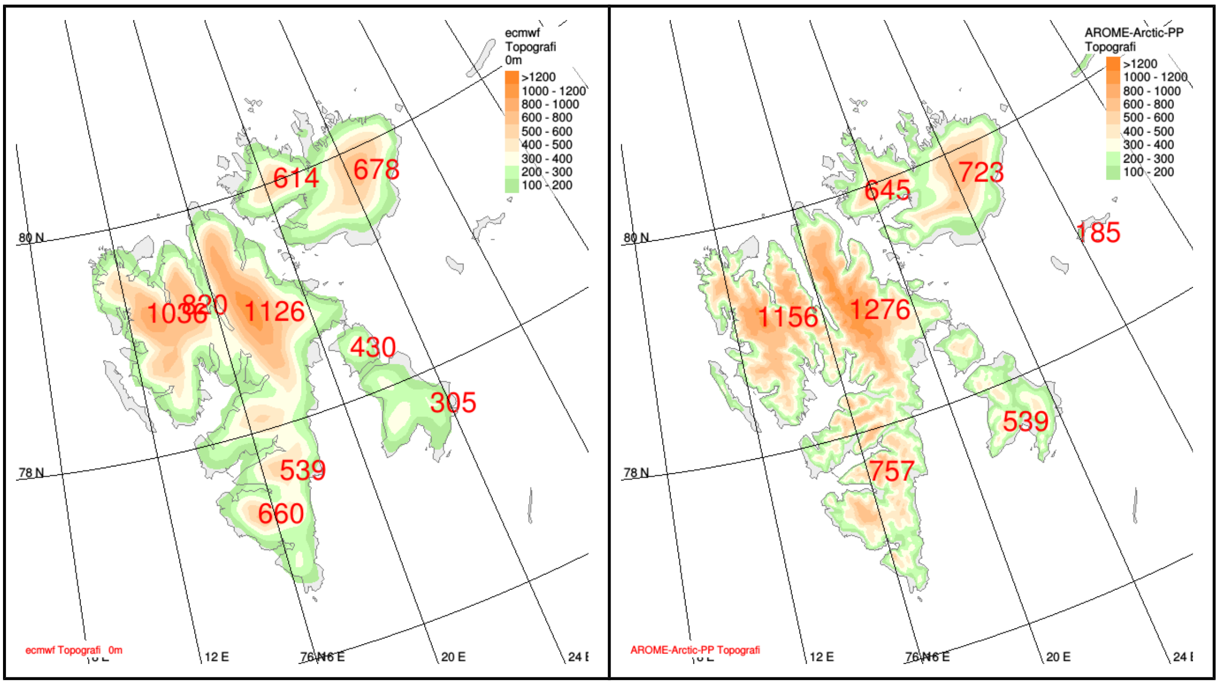

Monday 07 November 2016 00 UTC ecmf 1000 hPa Vertical integral of eastward water vapour flux
Mon 07 Nov 2016 00UTC @ECMWF VT: Mon 07 Nov 2016 00UTC - Tue 08 Nov 2016 00UTC 0-24h
Extreme forecast index and Shift of Tails (black contours 0,1,5,10,15) for: 2m mean temperature
-1
-0.9
-0.8
-0.7
-0.6
-0.5
0.5
0.6
0.7
0.8
0.9
1

Figure 3.1.1 Extreme forecast index (EFI) for vertically integrated water vapour flux (left) and 2-metre temperature (right) based on a 24-hour forecast from 7 November.



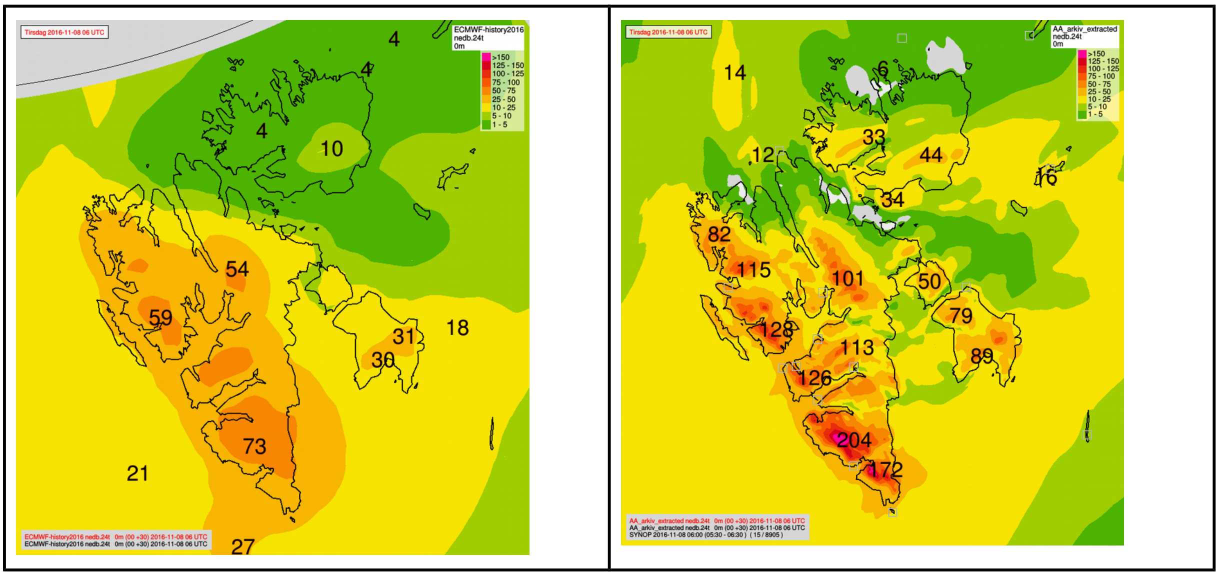

Tirsdag 2016-11-08 06 UTC
ECMWF-history2016
nedb.24t
0m
>150
125 - 150
100 - 125
75 - 100
50 - 75
25 - 50
10 - 25
5 - 10
1 - 5
4
4
4
10
54
59
31
30
18
73
21
27
ECMWF-history2016 nedb.24t 0m (00 +30) 2016-11-08 06 UTC
ECMWF-history2016 nedb.24t 0m (00 +30) 2016-11-08 06 UTC
Tirsdag 2016-11-08 06 UTC
AA_arkiv_extracted
nedb.24t
0m
>150
125 - 150
100 - 125
75 - 100
50 - 75
25 - 50
10 - 25
5 - 10
1 - 5
14
6
33
12
44
16
34
82
115
101
50
79
128
113
89
126
204
172
AA_arkiv_extracted nedb.24t 0m (00 +30) 2016-11-08 06 UTC
AA_arkiv_extracted nedb.24t 0m (00 +30) 2016-11-08 06 UTC
SYNOP 2016-11-08 06:00 (05:30 - 06:30) (15 / 8905)

Figure 3.1.2 Model topography on Svalbard for ECMWF HRES (left) and AROME Arctic (right) at the top and 24 hr accumulated precipitation from 7 November 2016 06 UTC to 8 November 2016 06 UTC in ECMWF HRES (left) and AROME Arctic (right) at the bottom.

3.2. Links with circulation patterns

Precipitation events as described above are often driven by large-scale flow patterns. The weather over the 'European' sector of the Arctic is influenced by the dominant weather pattern over the northern Atlantic. These weather patterns have been extensively studied in the literature, and are often referred to as Euro-Atlantic regimes, and the regime type and phase leads to strong anomalies of surface temperatures and precipitation over Europe (and Arctic). The variability of these regimes acts on many time-scales and some of this variability is predictable on monthly (Vitart (2014), Ferranti et. al. (2018)) and seasonal (Scaife et al., 2014) timescales. Hence this framework could also be useful for evaluation of the predictability in part of the Arctic.

There are several definitions of the Euro-Atlantic regime types. All of them include some form of the North-Atlantic Oscillation (NAO), which indicates the anomaly of the zonal wind over northern Atlantic. During its positive phase there is stronger than average westerly flow over northern Atlantic while in the negative phase the westerly winds are reduced or even reversed. Regime definitions often include indicators of the meridional modulation of the winds, such as blocking over northern Europe (often referred to as Scandinavian blocking). The regimes are often based on anomalies in the 500 hPa geopotential but can also be defined according to the position of the jet stream (Woolings et al. 2010).

In its classical form the NAO index (NAOCLASSIC) is measured by calculating the surface pressure difference between the Azores/Lisboa and Reykjavik, but it is today often based on the leading empirical orthogonal function (EOF1) for the variability over North Atlantic. The second EOF (EOF2) shows similarities with a blocking over Scandinavia. Operationally at ECMWF, 4 regimes are characterized using clustering techniques based on the work by Vautard (1990). In winter, these include positive (NAOP) and negative NAO (NAOM) phases, Scandinavian Blocking (BLOCK) and Atlantic Ridge (RIDGE). The corresponding spatial patterns are shown in Figure 3.2.1.

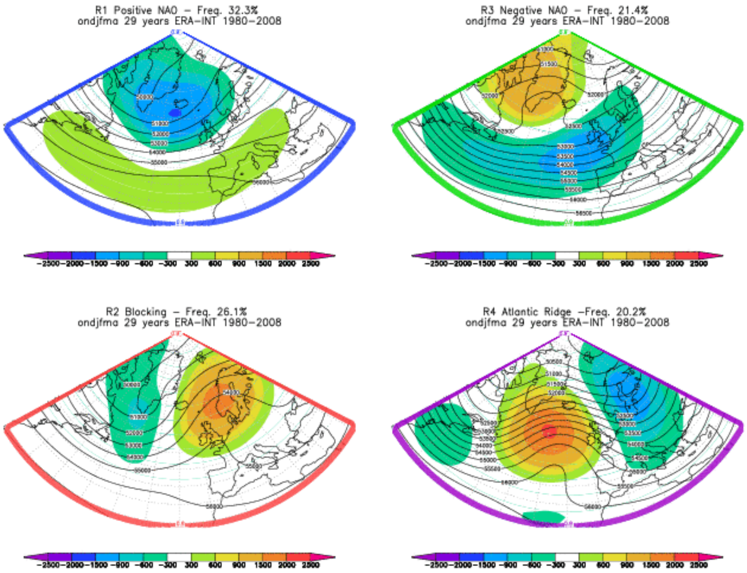

R1 Positive NAO – Freq. 32.3%
ondjfma 29 years ERA-INT 1980-2008
R3 Negative NAO – Freq. 21.4%
ondjfma 29 years ERA-INT 1980-2008
R2 Blocking – Freq. 26.1%
ondjfma 29 years ERA-INT 1980-2008
R4 Atlantic Ridge –Freq. 20.2%
ondjfma 29 years ERA-INT 1980-2008
-2500 -2000 -1500 -900 -600 -300 300 600 900 1500 2000 2500

Figure 3.2.1 Definitions of the 4 Euro-Atlantic regimes in terms of Z500 anomalies.

As mentioned above the Euro-Atlantic regimes also affect the occurrence of extremes in the Arctic. Figure 3.2.2 shows projections on the different regime definitions averaged over 20 extreme rainfall cases over Svalbard (rainfall events defined as days where at least 2 observation sites in Svalbard measured more than 10mm). For this type of extreme, the z500 anomaly of the days with extreme rainfall has the strongest positive projection onto the Scandinavian blocking pattern on average (Figure 3.2.2), similar to the findings of Serreze et al. (2015). It should be noted that for the case of November 2016, the projection was not onto this pattern, but rather a narrow and tilted ridge from Iceland towards northern Scandinavia.

To further study the relation between the Scandinavian Blocking pattern and anomalies in the Arctic, Figure 3.2.3 shows composites over 2-metre temperature (left) and precipitation (right) anomalies during days with a blocking over Scandinavia during DJF, based on 35 years of seasonal reforecasts with ECMWF System 4. On average the Scandinavian blocking is accompanied with warm temperatures over the north-eastern Atlantic and enhanced precipitation on Iceland, Northern Norway, Svalbard and eastern Greenland.

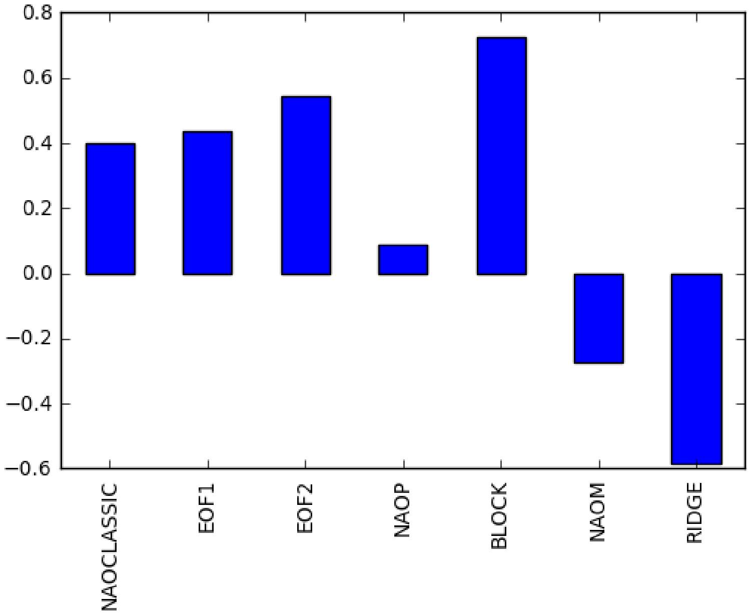


Figure 3.2.2 Average projection onto the 4 Euro-Atlantic regimes for the rainfall event over Svalbard. Negative values indicates projections on the reversed pattern.

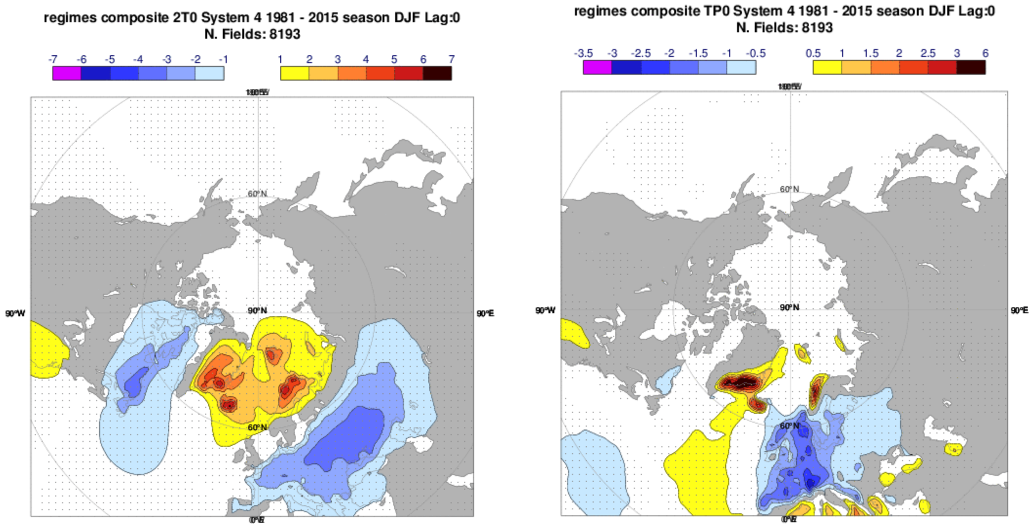


Figure 3.2.3 Composites of 2-metre temperature anomalies (left) and precipitation anomalies (right) during the Scandinavian blocking regime in System 4 for DJF 1981-2015.

Results from the previous figures suggest that Scandinavian blocking is associated (on average) to heavier than usual precipitation over Norway and Svalbard, a feature represented in the ECMWF seasonal forecasting System 4.

As a first approach to evaluate the seasonal re-forecasts discussed later (see section 5), we compute the blocking frequency in DJF for the November re-forecasts, using the method introduced by Tibaldi and Molteni (1990). Daily 500-hPa geopotential height fields for each ensemble member and over the 1993-2014 re-forecast period are used, and compared to corresponding data from ERA-Interim. Figure 3.2.4 shows the blocking frequency depending on the longitude for CNRM-CM6-1, SEAS5 and EC-Earth3 re-forecasts. All systems seem to capture reasonably well the variation of blocking frequency according to longitude, although model behaviour differs in the amplitude and exact position of the peaks in blocking activity. CNRM-CM6-1 (in red) clearly underestimates the blocking frequency over the Atlantic sector with respect to ERA-Interim, and overestimates blocking over the Pacific. SEAS5 captures quite well the Atlantic peak but underestimates blocking frequency in the Pacific sector, whereas EC-Earth3 finds a small shift in the peak of Atlantic blocking and overestimates the secondary peak around 60°E. Differences between SEAS5 and EC-Earth3 are larger than expected, given that the same atmospheric and ocean models (although with different versions) are used in both sets of re-forecasts. Note however that uncertainties in the ERA-Interim estimates cannot be excluded given the short re-forecast period, as shown by the more noisy curve than for re-forecasts which have larger samples due to the use of ensembles.

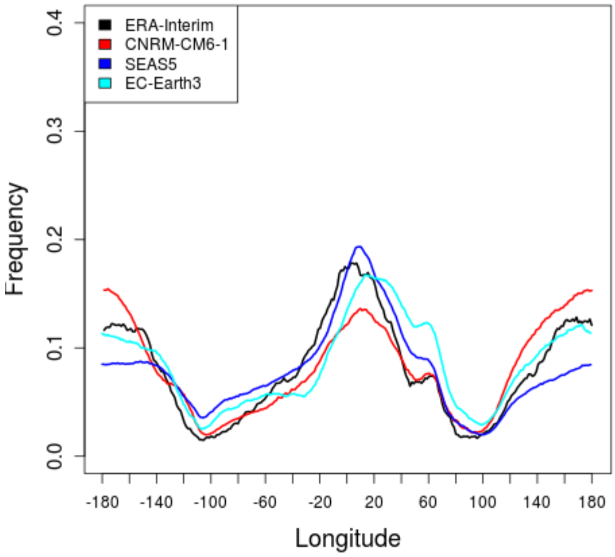


Figure 3.2.4. Frequency for Tibaldi and Molteni blocking index for DJF 1993-2014 in ERA-Interim (black), compared to seasonal re-forecasts with CNRM-CM6-1 (red), ECMWF SEAS5 (blue) and EC-Earth3 (light blue) for DJF initialized in November.

Beyond this evaluation of the mean statistics in the re-forecasts, actual skill in forecasting an increase in blocking activity is a necessary step for seasonal re-forecasts to be used as early indicators of possible heavy rainfall events over the region. Unfortunately, if current models show some skill in representing the NAO at a seasonal time scale (see section 5.1), little to no skill is found for blocking (not shown).

3.3. Predictability of short range forecasts

The short range forecasting capabilities for precipitation at Svalbard, exemplified with ECMWF HRES and the higher resolution limited area model AROME Arctic (hereafter AA) are evaluated in this section.

Verification statistics for daily precipitation for the period from March 2016 - April 2018 for 6 observation sites at Svalbard are presented in Figure 3.3.1 (location of observation sites given in Figure 3.3.2). Both ECMWF HRES and AA have a higher inherent spatial correlation than the observations, i.e. the models have less spatial variability than the observations. However, the higher horizontal resolution of AA adds value by more spatial variability. In general, the models overestimate the observed precipitation (exception: Ny-Ålesund), but AA forecasts less precipitation than ECMWF HRES (exception: Sveagruva). In particular, the precipitation amounts are overestimated at the three stations located in the interior fjords of Svalbard (Svalbard Airport, Adventdalen and Sveagruva). Opposite to this the three stations Ny-Ålesund, Hornsund and Isfjorden, situated closer to the west coast, experience less overestimation. The skill varies in space, e.g. the temporal correlation varies from 0.5-0.6 for Sveagruva to above 0.8 in Ny-Ålesund. A difference in skill is also seen between models where AA has lower MAE in 5 out of 6 stations and higher temporal correlation in 4 out of 6 stations than ECMWF HRES.

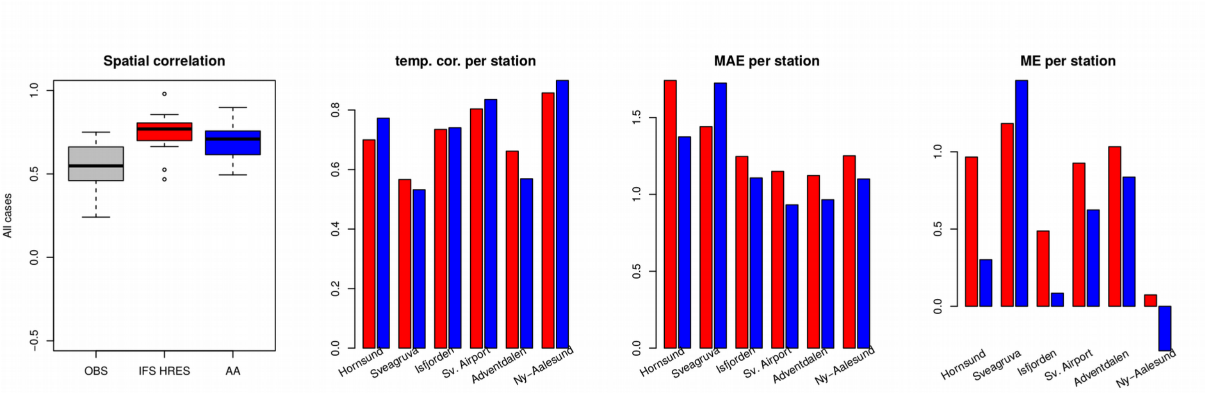


Figure 3.3.1 Verification statistics for daily precipitation on Svalbard for the period March 2016 to April 2018. To the left inherent spatial correlation between Svalbard stations for observations and forecasts, then temporal correlation, MAE and ME per station with ECMWF HRES (red) and AA (blue). All forecasts have lead time +30hr.

In the following we look at the forecast capabilities during severe precipitation events, which we define as a minimum of 2 (out of 6) observations > 10 mm/day. The days meeting this criteria are shown in Figure 3.3.4. Composites of forecasted precipitation from these cases are shown in Figure 3.3.2. The difference in model topography is a major driver of the differences seen between the models. The ECMWF HRES forecasts are smoother, with lower precipitation maxima compared to AA. The precipitation patterns (in particular from AA) indicate further that severe precipitation events at Svalbard are to a high degree steered by topographic effects. The maximum forecasted precipitation amounts, located at higher elevations, are at least twice what is seen at lower elevations. This also indicates that verification against observations is biased towards regions with minimum forecasted precipitation and sensitive to the representativeness of the nearest grid point to the observation sites.

Despite local differences between the composites from ECMWF HRES and AA, the average values for the plotted area in Figure 3.3.2 are very similar with 6.2 and 6.1mm/day for ECMWF HRES and AA, respectively. On average, the potential added value of AA is therefore related to a spatial redistribution of the precipitation. However, in some situations there is (up to 36%) more precipitation in ECMWF HRES (top row Figure 3.3.3), while there are other situations with (up to 22%) more precipitation in the AA forecast (bottom row Figure 3.3.3). A tendency for ECMWF HRES to produce more precipitation than AA when parts of the air flow are from the north is seen. In the opposite situations, AA produces more precipitation when the air flow is mainly from the south. The results indicate that the severe Svalbard precipitation events are sensitive to both the moisture transport towards Svalbard and to the representation of the topography.

Various verification plots for the severe precipitation events are presented in Figure 3.3.4. The time series show the highest observed and forecasted (from the observation sites) daily precipitation values at Svalbard. A relatively good day-to-day agreement between forecasted and observed maximum precipitation is found with a correlation of 0.84 for AA and 0.76 for ECMWF HRES. However, both models fail to forecast the highest observed precipitation amounts, but AA are closer to the observations (e.g. November 2016, September 2017 and January 2018). The highest precipitation amounts at Svalbard are observed at Ny-Ålesund which therefore dominate the presented time series. However, Ny-Ålesund is the only Svalbard stations where AA adds value regarding simulating the precipitation peak events (top right part of Figure 3.3.4). Furthermore, the temporal correlation, MAE and bias station by station indicate a shift towards less added value of AA in the high precipitation situations. In addition ECMWF HRES are in better agreement with the observed spatial correlation than AA. It can therefore be argued that to some extent the added value of AA compared to ECMWF HRES is less, or more complex to extract, for the severe events than when all cases are included.

Also for the severe precipitation events the forecasts overestimate the precipitation at Svalbard Airport, Adventdalen and Sveagruva. Furthermore, the models are not able to predict the local differences between Svalbard Airport and Adventdalen. The same behaviour is also seen for ECMWF HRES and AA in a Year of Polar Prediction Special Observing Period 1 (YOPP-SOP1) model intercomparison (Køltzow et al., 2018). However, in another set-up of AROME provided by Meteo-France, also with 2.5 km horizontal grid spacing, the lower observed precipitation amounts in Adventdalen are better captured. This indicates that 2.5 km horizontal grid spacing has the potential to forecast very small scale precipitation patterns, but there is also some sensitivity to the configuration of the model. It should also be mentioned that there are some uncertainties with respect to precipitation observations at Svalbard airport and Adventsdalen which are under investigation.

Many of the severe precipitation events on Svalbard are rain on snow events, i.e. rain on snow covered surfaces and a later re-freeze with implications for infrastructure and wildlife (Hansen et al., 2014). Correct forecasts of the precipitation phase are therefore important. Since direct observations of the precipitation phase are rare in time and space we use 2m air temperature as a proxy for the precipitation phase. Averaged over the severe precipitation events the forecasts have a negative temperature bias (-1.03C for AA and -1.55C for ECMWF HRES) indicating too much solid precipitation and too little rain. We further assume a temperature threshold of +1C to distinguish between rain and solid precipitation. This threshold results in 73% of the AA and 59% of the ECMWF HRES precipitation being forecasted as rain. If we instead use the observed temperatures the numbers are 76 and 73%, respectively. The results indicate that AA does a reasonable good job on the precipitation phase and adds value on this aspect compared to ECMWF HRES.

In summary, the higher resolution limited area model AA adds value to the global ECMWF HRES in several aspects regarding forecasting precipitation at Svalbard. However, the added value is not evenly distributed (e.g. in space, time, on parameters or specific precipitation characteristics) and there are, even with 2,5 km horizontal resolution and the present configuration of AA, weaknesses in the predictability of severe precipitation events at Svalbard.

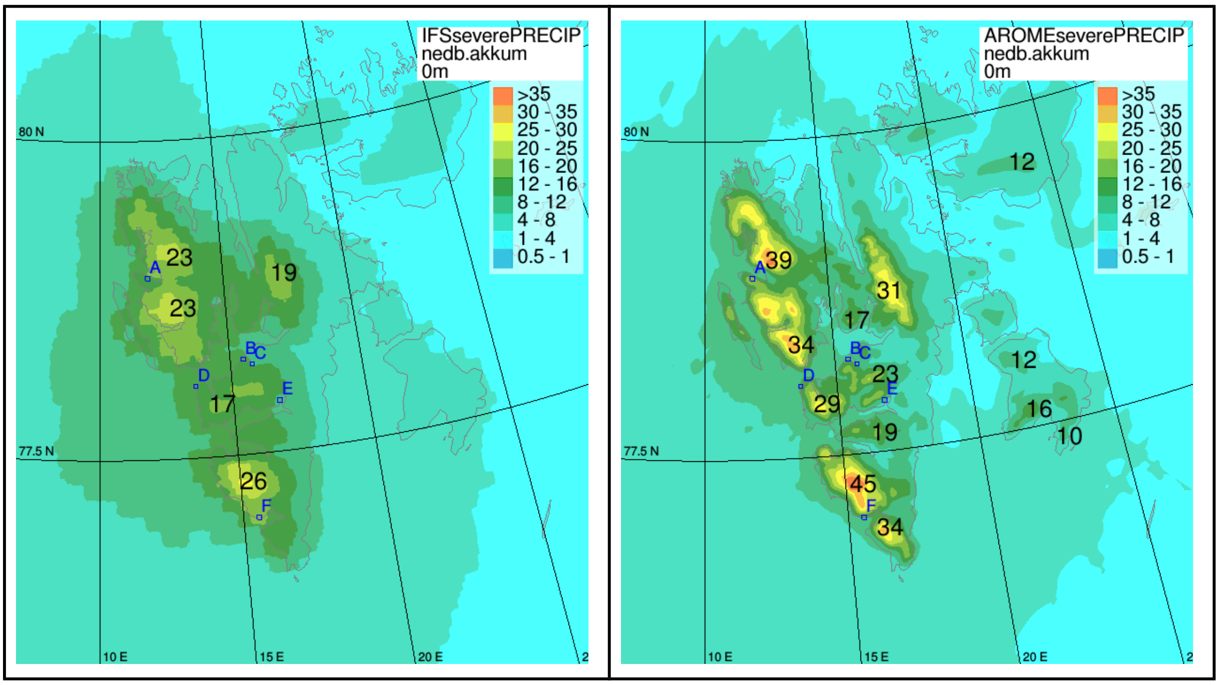
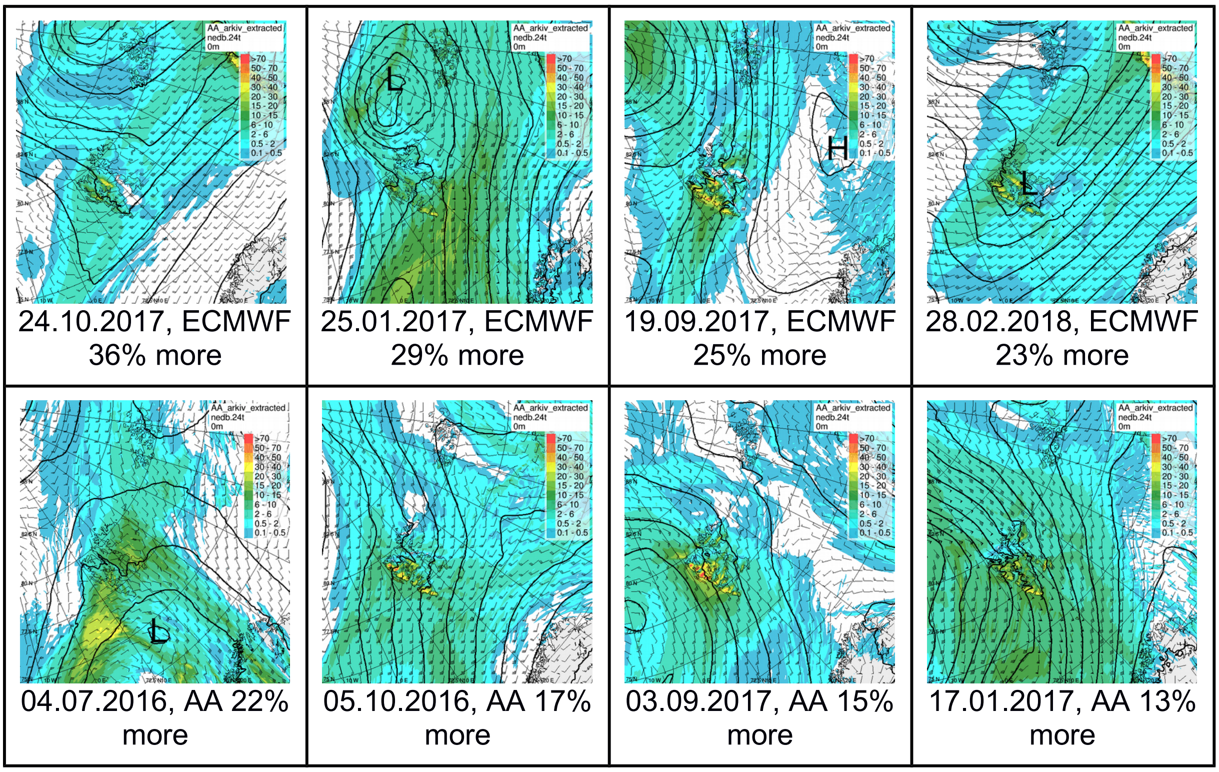


Figure 3.3.2. Daily forecasted precipitation averaged over severe precipitation cases. Observation sites used in verification statistics; A - Ny-Ålesund, B - Svalbard Airport, C - Adventdalen, D - Isfjorden radio, E - Sveagruva and F - Hornsund. Numbers in black are maximia averaged over all forecasts.



24.10.2017, ECMWF 36% more

25.01.2017, ECMWF 29% more

19.09.2017, ECMWF 25% more

28.02.2018, ECMWF 23% more

04.07.2016, AA 22% more

05.10.2016, AA 17% more

03.09.2017, AA 15% more

17.01.2017, AA 13% more

Figure 3.3.3. From AA forecasts, daily accumulated precipitation, MSLP and 925h Pa wind from time in the middle of the precipitation accumulation period. In the top row; the 4 events where ECMWF HRES forecast the highest precipitation amounts compared to AA (averaged over the plotted region), in the bottom row; the 4 events where AA forecast the highest precipitation amounts compared to ECMWF HRES.

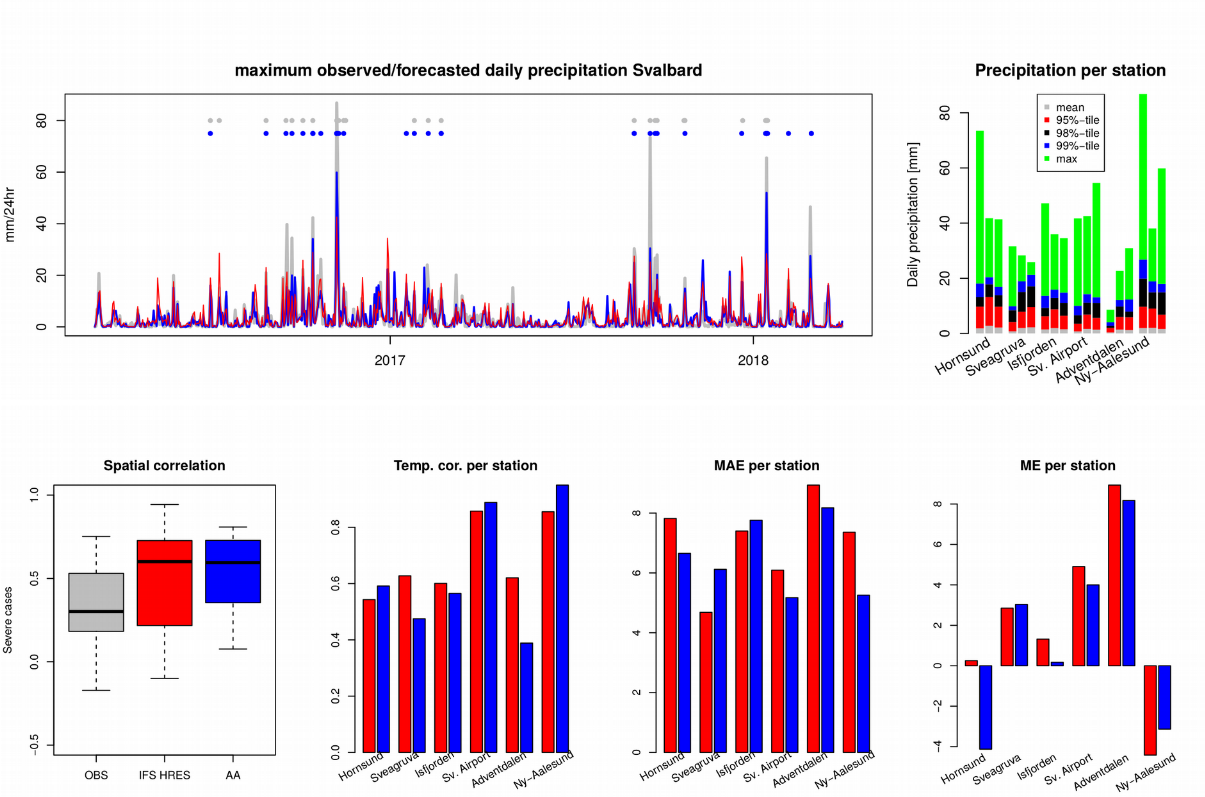


Figure 3.3.4 Top left; time series of maximum daily precipitation (observed and forecasted) at Svalbard observation sites from March 2016 to April 2018. Top right; for 6 stations the mean value (grey), the 95, 98 and 99%-tile (red, black and blue, respectively) and the max value (green) for 24h observed precipitation (first column), ECMWF HRES (second column) and AA (last column). Second row; statistics over severe precipitation events marked with blue dots in the time series. To the left inherent spatial correlation between Svalbard stations for observations and forecasts, then correlation, MAE and ME per station with ECMWF HRES (red) and AA (blue). All forecast has lead time +30hr. Location of stations are given in Figure 3.3.2.

4. EVALUATION OF SHORT AND MEDIUM-RANGE FORECASTS OVER THE ARCTIC

4.1. Evaluation of medium-range forecasts over the Arctic

This section aims to document the medium-range forecast skill over the Arctic and contrast it to the predictability over the full northern hemisphere. The verification in this section is based on ECMWF operational global medium-range forecasts from DJF 2016-2017 and JJA 2017. The ECMWF medium-range forecasts include a high-resolution forecast with horizontal resolution of 9 km (HRES) and a 51 member ensemble with 18 km resolution (ENS). The verification here is mainly made against the operational analysis. Albeit the disadvantages of using an analysis as verification, the uneven distribution of observations in the Arctic would make the results difficult to interpret. Work package 4 in Applicate will assess the validity of the analysis. We will also use SYNOP observations to evaluate biases in weather parameters. The scores presented here are the spatial anomaly correlation coefficient (ACC), root-mean-square error (RMSE) and mean error (ME).

Figure 4.1.1 shows the evolution of the skill in terms of ACC for z500 over the Arctic (65N-90N) and Northern Hemisphere (20N-90N) since 1990 for HRES. The plot also includes results for forecasts based on ERA-Interim (Dee et al. 2011). ERA-Interim used a fixed forecast system during the whole period and can therefore be used to determine the natural variability in the predictability. The general result is a steady improvement in the scores over the decades, with similar pace for the Arctic and the Northern Hemisphere. The factors behind the long-term improvements are discussed in Magnusson and Källén (2013). The scores have been somewhat lower for the Arctic than over the whole hemisphere during most of the period. In the early 90s the ACC for HRES was similar for both regions. However, looking at the score for ERA-Interim, the Arctic seems to have been relatively predictable during that period (1990-1994). The same holds true for the last year, where the relative improvement over the Arctic could be attributed to natural variability.

Figure 4.1.2 shows the spatial distribution of RMSE for geopotential at 200, 500 and 850 hPa in 5-day forecasts verified against the operational analysis. The 200 hPa level shows low errors over the central Arctic. This could be explained by the fact that the 200 hPa level is in the stratosphere over the Arctic, which is dominated by slower time-scales than in the troposphere. It is mainly in the lower troposphere that errors are larger over the Arctic than in the mid-latitudes. In winter-time high errors are especially found over the north Atlantic where the storm track leads to high variability. However, during summer the largest RMSE on 500 and 850 hPa level is found in the central Arctic. It is a sign of the Arctic being more dynamically active during the summers, and therefore more unpredictable.

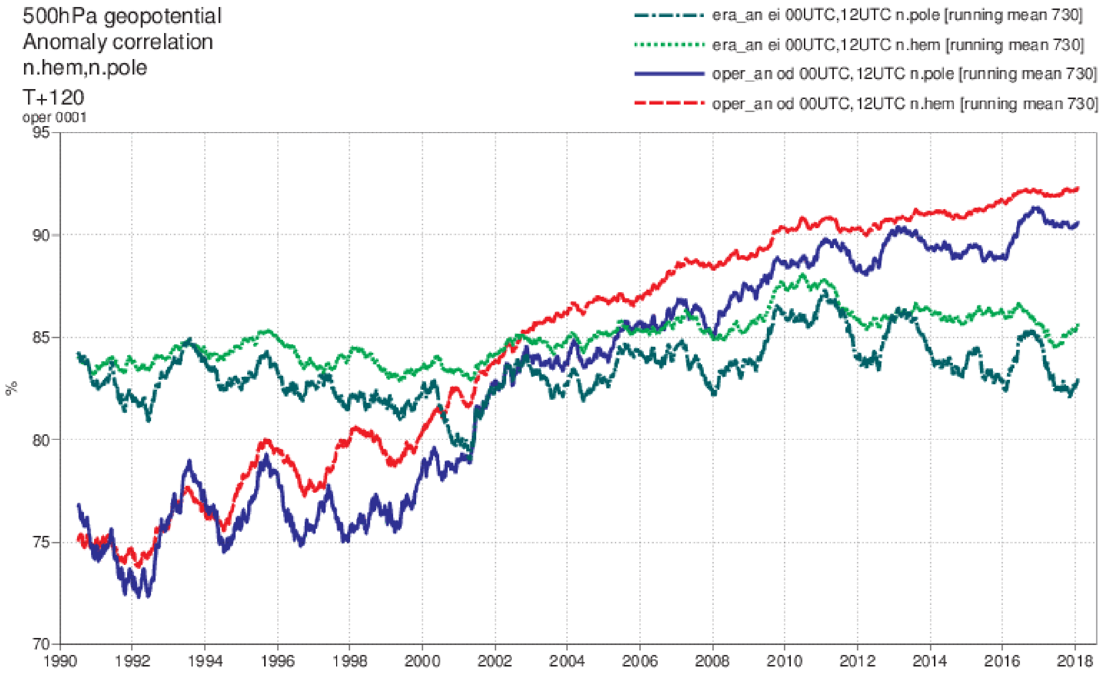


Figure 4.1.1 Evolution of day 5 ACC in ECMWF for N.Hem (red) and Arctic (blue) and ERA-Interim N.Hem (light-green) and Arctic (dark-green).

Figure 4.1.3 shows maps of ME for HRES forecasts verified against SYNOP observations The results are for 24-hour minimum (top) and maximum (bottom) temperature. The daily extremes have been determined from hourly observations and forecast values spanning 24 to 48 hours into the forecast. The scores are presented both for DJF (left) and JJA (right). The same type of verification is available for 10-metre wind speed, total cloud cover and 24-hour precipitation but is not presented here.

On the land areas over the Arctic, a positive bias is dominating for minimum temperature in DJF. This bias is mainly occurring during nights with strong surface inversions, a common issue for weather and climate models. For JJA, the largest bias appear in the maximum temperature with a cold bias.

In the next section we are going to look further into errors in “weather parameters”, and compare the global model with the regional Arome Arctic model.

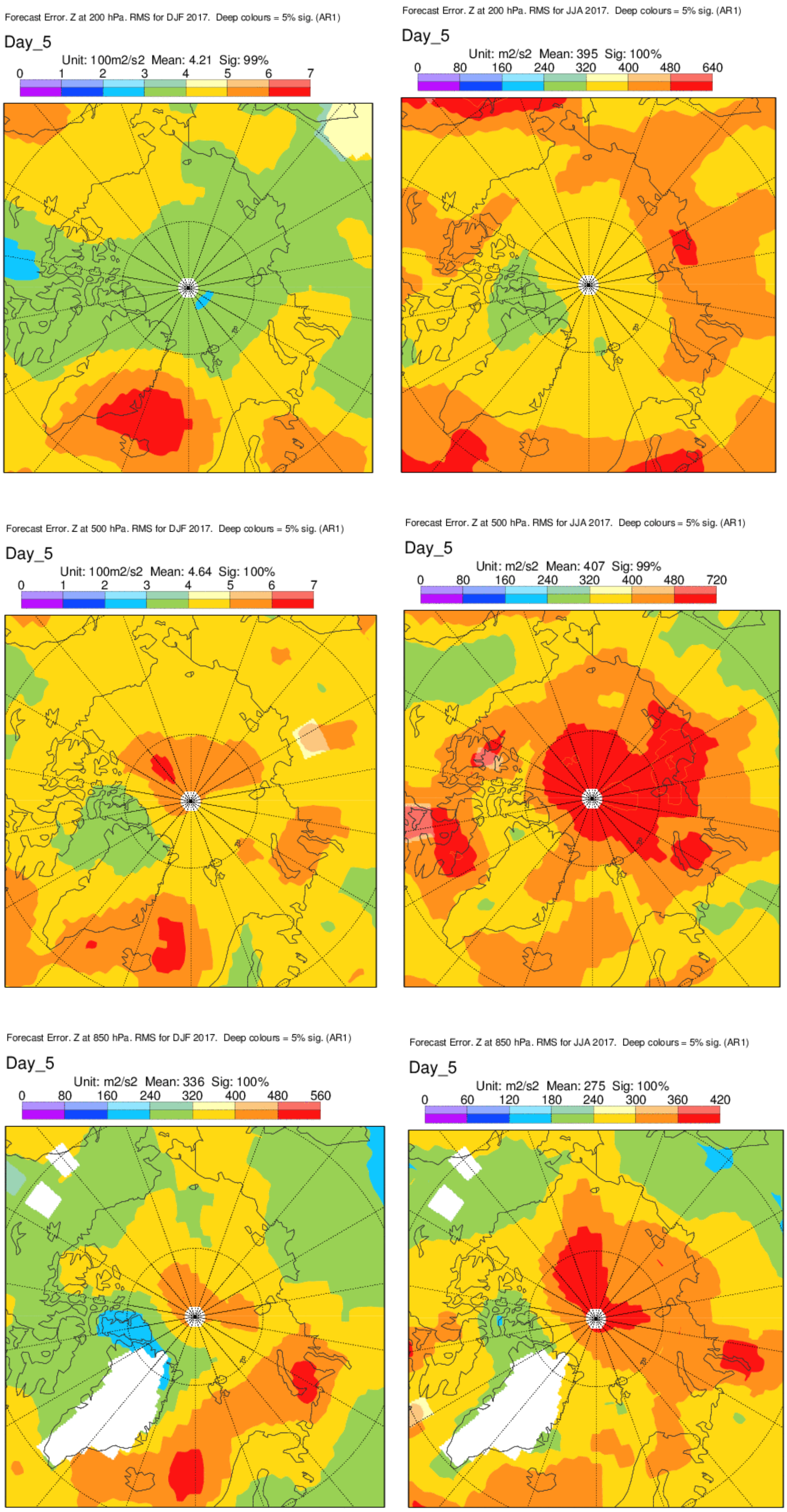

Forecast Error. Z at 200 hPa. RMS for DJF 2017. Deep colours = 5% sig. (AR1)
Day_5
Unit: 100m2/s2 Mean: 4.21 Sig: 99%
0 1 2 3 4 5 6 7
Forecast Error. Z at 200 hPa. RMS for JJA 2017. Deep colours = 5% sig. (AR1)
Day_5
Unit: m2/s2 Mean: 395 Sig: 100%
0 80 160 240 320 400 480 640
Forecast Error. Z at 500 hPa. RMS for DJF 2017. Deep colours = 5% sig. (AR1)
Day_5
Unit: 100m2/s2 Mean: 4.64 Sig: 100%
0 1 2 3 4 5 6 7
Forecast Error. Z at 500 hPa. RMS for JJA 2017. Deep colours = 5% sig. (AR1)
Day_5
Unit: m2/s2 Mean: 407 Sig: 99%
0 80 160 240 320 400 480 720
Forecast Error. Z at 850 hPa. RMS for DJF 2017. Deep colours = 5% sig. (AR1)
Day_5
Unit: m2/s2 Mean: 336 Sig: 100%
0 80 160 240 320 400 480 560
Forecast Error. Z at 850 hPa. RMS for JJA 2017. Deep colours = 5% sig. (AR1)
Day_5
Unit: m2/s2 Mean: 275 Sig: 100%
0 60 120 180 240 300 360 420

Figure 4.1.2. Day 5 RMSE for geopotential for 200 hPa (top), 500hPa (middle) and 850 hPa (bottom) for DJF 2016-2017 (left) and JJA 2017 (right).

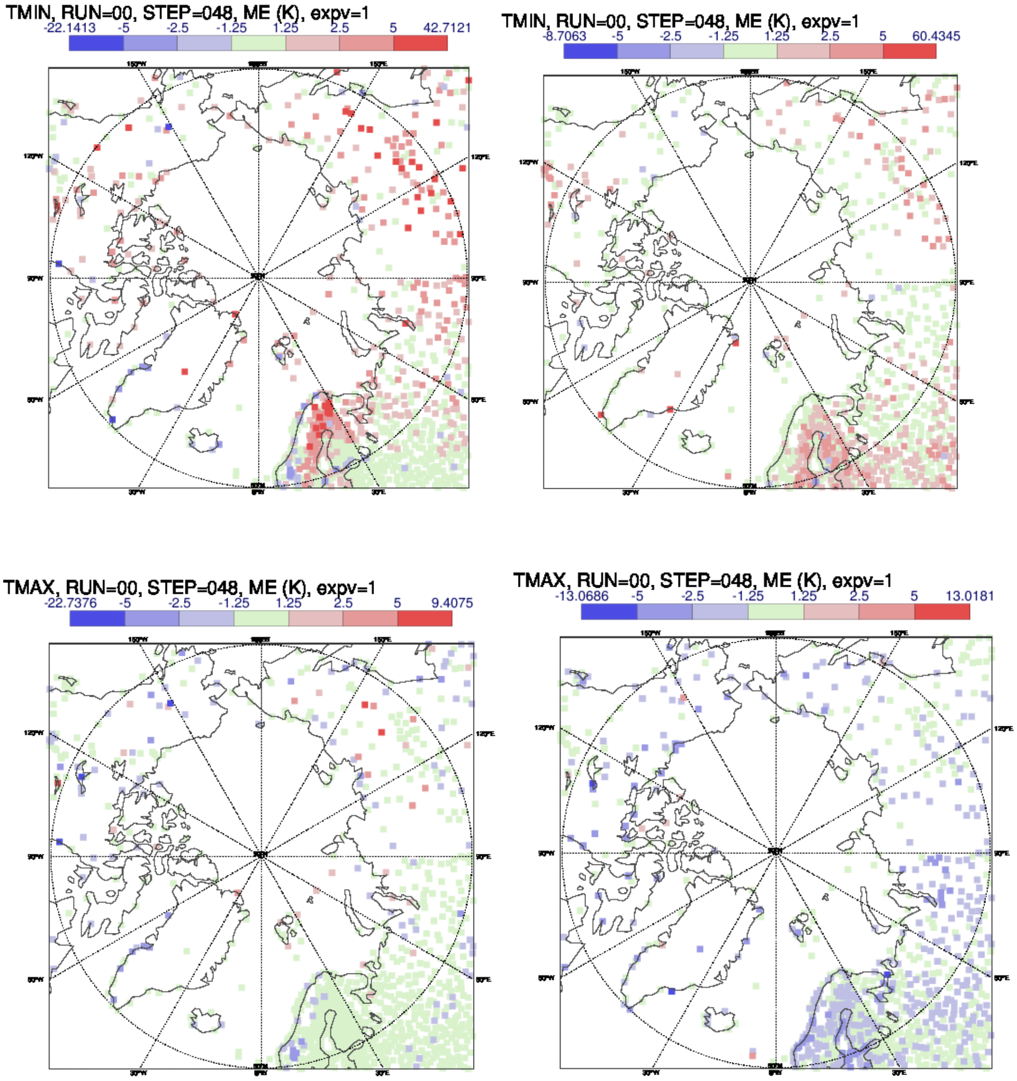


Figure 4.1.3. Day 2 mean error of daily minimum (top) and maximum (bottom) temperature for DJF 2016-2017 (left) and JJA 2017 (right).

4.2. Comparison of short range forecasts with observations

Limited area models can, for targeted areas, be compared with global models, and have shown improved forecast skill associated with the use of optimized physics and finer horizontal and vertical resolution (Jung et al., 2016). In the following we compare forecasts from the high resolution limited area model AROME Arctic, hereafter AA and the ECMWF HRES forecasts (see section 2.1 for model descriptions).

The short range operational forecasts (+9, +12, +15, …, +30h) from ECMWF HRES and AA are compared over the period 10 March 2016 - 31 April 2018. The models are compared against quality controlled Norwegian observations derived from the MET Norway observational database (eklima.met.no). The comparison is divided into 6 roughly homogeneous regions based on the knowledge and experience of operational forecasters at MET Norway (Figure 4.2.1). The different regions make the comparison more useful as model differences and errors are more easy to interpret. It is the raw model output that is compared and no post-processing, spatial methods or other adjustments are employed to models or observations. Operational systems are in a continuous process of improvement and several model upgrades for both systems were introduced in the inter-comparison period, but the grid spacing of both models was kept constant.

In the comparison we use Mean Absolute Error (MAE), Mean Error (ME) and the categorical Equitable Threat Score (ETS) and Frequency Bias (FB) described in section 2.2.1. This selection of metrics does not cover all aspects of the forecast quality, but gives a good overview of some main characteristics and differences between the model systems. The comparison of AA and ECMWF HRES is presented in Figure 4.2.2 and 4.2.3 and the main features are summarized below.

Mean Sea Level Pressure (MSLP) forecasts have small errors, but ECMWF HRES scores consistently better than AA indicating a better description of the large scale systems. This is a behavior often seen in the comparison of operational global and regional models. Some possible mechanisms to explain this are e.g. better assimilation of large scale weather in global models, inaccurate treatment of lateral boundary data in regional models, slightly different “tuning” of global (focus: low pressure development) and regional (focus: 10m wind speed) models and more small scale noise with higher resolution. Mountain areas have the most pronounced errors with a substantial systematic part for both model systems. Furthermore, the errors are larger and more systematic in winter. How observed and forecasted surface pressure are reduced to MSLP might explain part of the more pronounced mountain and winter errors (Pauley, 1998).

2m air temperature (t2m) errors vary between regions. Smaller errors are seen when the influence from the relatively well represented sea surface temperatures is significant (islands, coast) and larger errors are present in e.g. fjords, inland and mountains. As discussed in Section 4.1, ECWMF has a warm bias in cold conditions (night-time), and a cold bias for daily maximum in summer-time. AA has lower MAE than ECMWF HRES with the exception of islands. The difference in MAE is particularly pronounced in complex terrain (fjords, inland and in mountain areas). A substantial part of the differences can be attributed to higher systematic errors in ECMWF HRES. AA benefits from the higher horizontal resolution with a better description of complex terrain, coast lines and other local heterogeneities feeding local differences in temperature. A simple height correction of ECMWF HRES reduces some of these errors in the presence of complex terrain (not shown). Largest t2m errors are present in winter and are often connected to difficulties in representing the stable boundary layer properly.

10m wind speed (S10m) errors measured with MAE and ME increase for regions exposed to high wind speeds. Regionally this is seen from inland to fjords, coast/Svalbard and mountains, while seasonally the errors increase from summer to spring/autumn and winter. In general, AA has smaller errors than ECMWF HRES (excepted over islands and inland). Both model systems have regions with pronounced systematic errors, but these vary between the systems. In general AA forecasts stronger winds than ECMWF HRES. ECMWF HRES has a larger negative bias than AA in the mountains, Svalbard, coast and fjords, but AA overestimates the wind speed for inland and islands. The categorical scores show that the forecast skill decreases with increasing wind speed thresholds, but AA performs better than ECMWF HRES for all chosen thresholds. Part of this difference is due to a better frequency climatology for AA, e.g. for more than 20.8m/s AA has a FB of 0.4-0.7 (depending on season) while ECMWF HRES very rarely forecast this (FB < 0.1).

2m air relative humidity (rh2m) errors vary between regions and seasons. Smallest MAE is found for islands (ECMWF HRES slightly higher (systematic) errors in the cold season). Opposite to this AA has higher (systematic) errors inland in spring (and winter/summer). An improved physical description of the surface in AA, which is under implementation, reduces this error. In general the most pronounced errors are found in mountain areas, but with a similar forecast quality from both models.

Total cloud cover (TCC) has lower MAE for ECMWF HRES for all regions and seasons. Norwegian TCC observations are manual observations and valid for a large spatial area that better fits the ECMWF HRES resolution. A simple smoothing of the TCC fields from AA to the resolution of ECMWF HRES reduce parts of the difference in skill (not shown). For the systematic errors it is difficult to see specific patterns in model system behaviour.

24h accumulated precipitation (P24) errors measured with MAE are in general slightly higher for ECMWF HRES than AA. At least part of this difference is due to an overestimation in ECMWF HRES. For both model systems the highest MAE is found for coast, fjords and mountain stations. Notice that due to the undercatch of solid precipitation in precipitation gauges (not adjusted for in this comparison) it is very difficult to draw conclusions for seasons that are dominated by solid precipitation. Most likely, both model systems underestimate the solid precipitation (Køltzow et al., 2018). With respect to categorical scores the forecast skill decreases with increasing thresholds. For the lowest thresholds (e.g. precipitation / no precipitation) AA has higher ETS and better FB than ECMWF HRES. For the highest thresholds AA has a better FB (ECMWF HRES rarely forecast the highest thresholds), but this is not reflected in a higher skill measured in form of ETS. The latter might be due to the “double penalty issue” that is a well known issue for high resolution models (Mass et al., 2002).

In summary, the higher resolution of AA adds value because of a better description of small scale forcings (e.g. complex topography, coast lines, surface heterogeneities). However, the added value varies and is not necessarily seen for all parameters, regions or seasons. The results presented in this inter-comparison are in agreement with the findings in a model-intercomparison during the YOPP-SOP1 which in addition to AA and ECMWF HRES also include the high resolution model systems Canadian Arctic Prediction System, CAPS, and an AROME version based on the Meteo France set-up (Køltzow et al., 2018). Køltzow et al. (2018), which is also work done in the APPLICATE project, go in more detail on several aspects of the Arctic forecast skill assessment which is beyond the scope for this report.

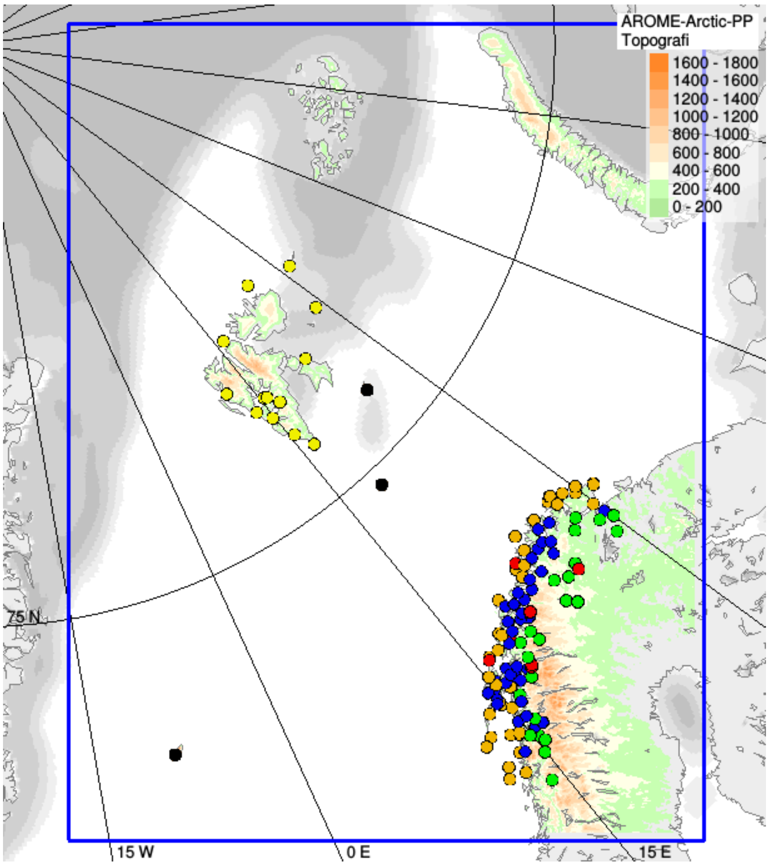


Figure 4.2.1 The integration domain for AA and Norwegian SYNOP observation used for comparison are plotted as black (Islands), yellow (Svalbard), orange (coast), blue (fjords), green (inland) and red (mountains) circles. In grey colours sea ice concentration from ECMWF HRES 1 March 00 UTC 2018 and in green/brown colours the model topography from AROME Arctic.

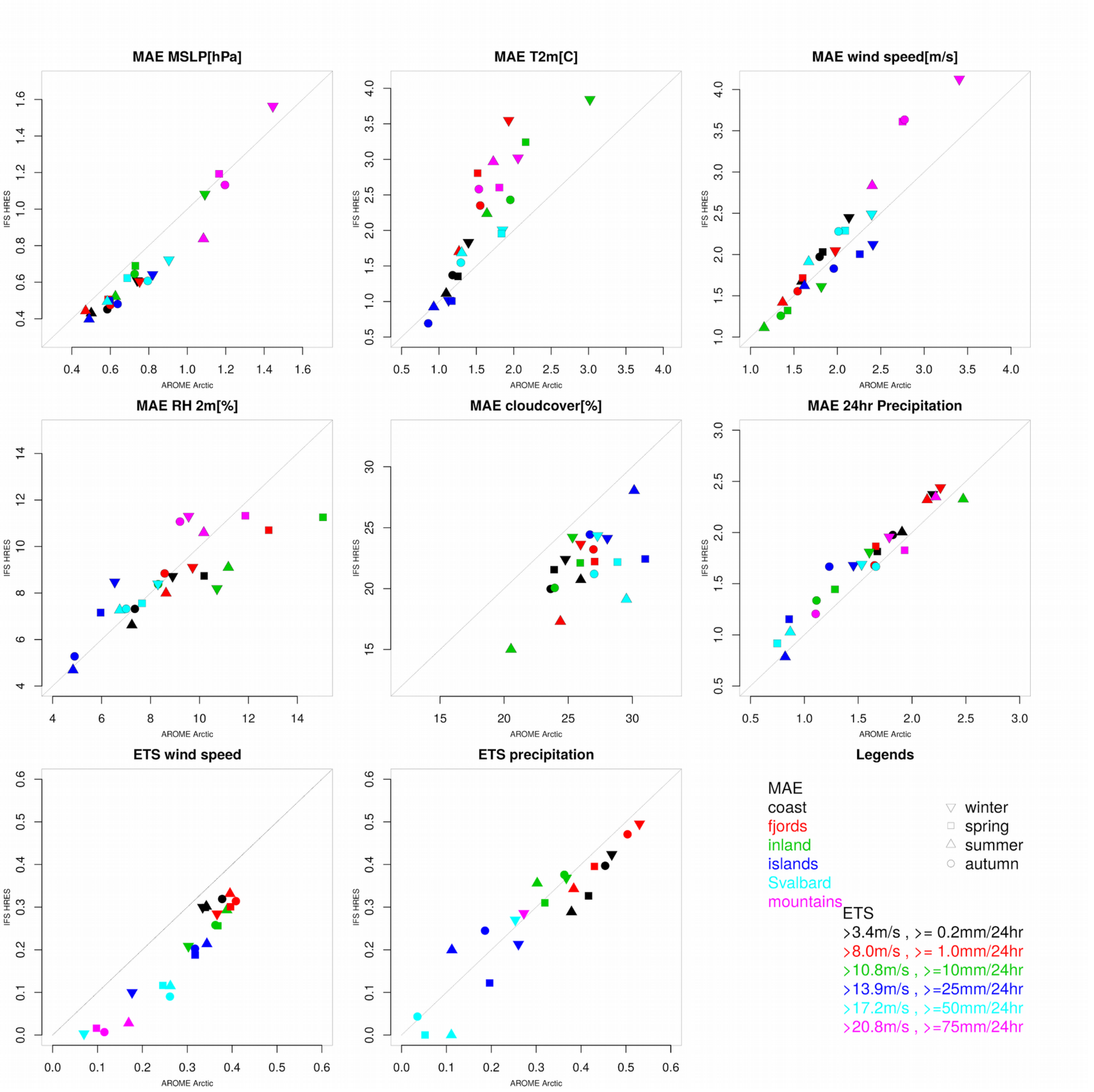


Figure 4.2.2 Comparison of forecast skill between AROME Arctic (x-axis) and ECMWF HRES (y-axis) for March 2016 to April 2018. Skill is computed against all Norwegian SYNOP stations available in the common domain (in total ~120 stations, see Figure 4.2.1, but not all measure all parameters). The 6 upper panels show Mean Absolute Error for parameters Mean Sea Level Pressure, 2m air temperature and relative humidity, 10m wind speed, total cloud cover and daily precipitation for different seasons (shape of plot) and regions (color of plot). The two lower panels show Equitable Threat Score (ETS) for different seasons (by shape) and exceeding thresholds (by color) in 10m wind speed and daily precipitation for all stations (too little data for some thresholds to get robust results if divided in regions). Lead times included are +9, +12, +15, +18, +21, +24, +27 and +30h.

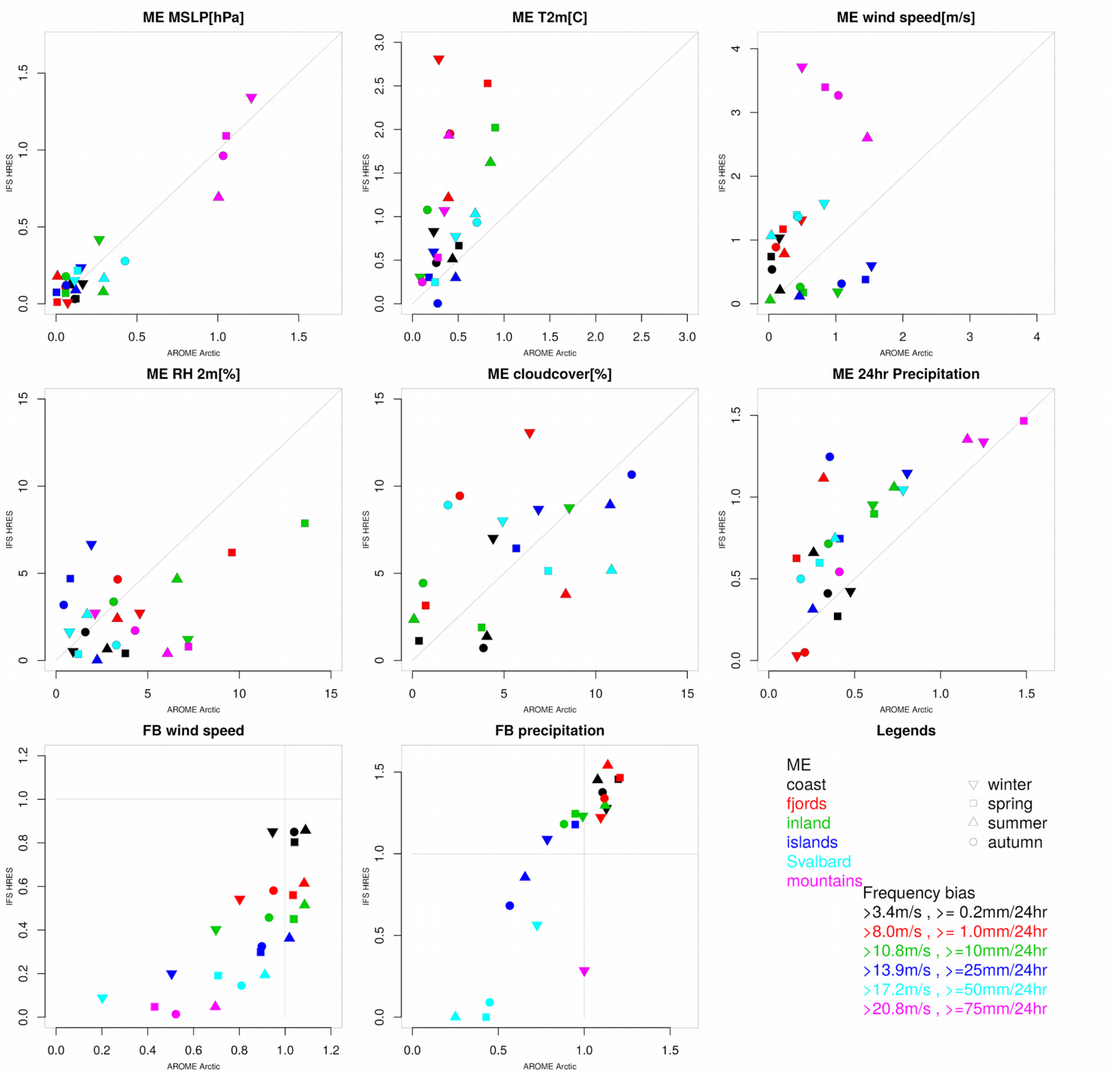


Figure 4.2.3 Comparison of forecast skill between AROME Arctic (x-axis) and ECMWF HRES (y-axis) for March 2016 to April 2018. Skill is computed against all Norwegian SYNOP stations available in the common domain (in total ~120 stations, see Figure 4.2.1, but not all measure all parameters). The 6 upper panels show the absolute value of the Mean Error for parameters Mean Sea Level Pressure, 2m air temperature and relative humidity, 10m wind speed, total cloud cover and daily precipitation for different seasons (shape of plot) and regions (color of plot). The two lower panels show Frequency Bias (FB) for different seasons (by shape) and exceeding thresholds (by color) in 10m wind speed and daily precipitation for all stations (too little data for some thresholds to get robust results if divided in regions). Lead times included are +9, +12, +15, +18, +21, +24, +27 and +30h.

5. SEASONAL FORECAST QUALITY OVER THE NORTHERN HEMISPHERE MIDLATITUDES AND ARCTIC

5.1. Atmospheric predictability for boreal winter

This section describes the ability and deficiencies of current state-of-the-art seasonal forecasting systems in reproducing northern hemisphere mid-latitude atmospheric variability for common climate variables (i.e. surface air temperature, sea level pressure and precipitation). The analysis is focused on coupled seasonal re-forecasts from the stream 1 of APPLICATE WP5 experiments (see Deliverable 5.1). The reference data for the analysis of sea level pressure and surface air temperature is ERA-Interim, while for precipitation the reference is GPCP V2.2. All results in this section were produced after simple bias correction.

Based on evaluations of the Fair Continuous Ranked Probability Skill Score, the winter re-forecasts of surface air temperature (Fig 5.1.1) show consistent skill across models in most of the Pacific Ocean, Northeast Atlantic and Barents Sea, being the only Arctic region with consistent skillful predictions. The surface temperature predictability in the Barents Sea is most likely related to the ocean/sea ice state. Interestingly, all models fail to predict temperature around Iceland showing negative values in the FCRPSS. In terms of sea level pressure (Fig 5.1.2) all models show consistent skillful prediction over the Aleutian region, tropical western Pacific, Pacific and Gulf coasts of North America, and South Asia. Skill in precipitation forecasts is patchy and low for all models, reduced to parts of southern North America and East China (Fig. 5.1.3). FCRPSS shows consistently lower scores than a typical deterministic score like anomaly correlation coefficient. To illustrate this, we show maps of sea level pressure anomaly correlation between models and ERA-Interim in Fig 5.1.4. Comparing Fig 5.1.2 with Fig 5.1.4 shows large differences depending on the metric used. FCRPSS is much lower because it does not detect skill that is potentially achievable with the mean of a large ensemble if the signal to noise ratio is too small in the models as opposed to anomaly correlation.

Large scale climatic variability such as the El Niño Southern Oscillation (ENSO) or the North Atlantic Oscillation (NAO) is associated with climatic teleconnection patterns extending over large parts of the planet (e.g. global or hemispheric). Skillful prediction of such climatic variability would translate in widespread predictability over large areas. All models show skillful prediction of ENSO from November to April in the winter forecasts (Fig 5.1.5) with anomaly correlation coefficients of over 0.8 up to five months lead time. Contrasting with ENSO, the winter NAO skill is considerably lower (Fig 5.1.6). However, the low r-values are in part due to small ensembles: for instance CNRM-CM6 and GloSea5 have statistically significant r-values of 0.51 and 0.43 with thirty members and fourteen members, respectively. Something similar happens with the multi-model ensemble, which has a statistically significant r-value of 0.5. The ratio of predictable components (RPC) in Fig 5.1.6 indicates that all 10-member individual model ensembles are overconfident (CNRM-CM6 only slightly), while the multi-model ensemble is underconfident. The latter is consistent with results using larger ensembles which show that the signal to noise ratio is too small in many models compared to observations (Baker et al 2018). Although this highlights a model deficiency, it also suggests that skilful forecasts of the NAO can be made by taking the mean of a large ensemble to extract the predictable signal.

To help interpret the predictive capacity of the systems, linear correlations of winter (DJF) surface air temperature, sea level pressure and precipitation with the NINO3.4 index (Fig 5.1.7) and the NAO index (Fig 5.1.8) were done using a 200 year control preindustrial simulation with EC-Earth3.2. The spatial similarity between ENSO teleconnections (positive and negative correlations) and the improved forecast skill in the climate variables displayed in Figs 5.1.1-3 suggests that ENSO is the main source of predictability on seasonal timescales for the extratropical northern hemisphere - a feature common to seasonal prediction systems for many years (Doblas-Reyes et al., 2013). For a given model, as opposed to ENSO, the low/intermediate predictive skill of the NAO may hinder its capacity to provide skilfull predictions in the NAO influence regions through the teleconnections displayed in Fig 5.1.8. However, we note that higher NAO skill is potentially available with a larger ensemble (Athanasiadis et al 2017).

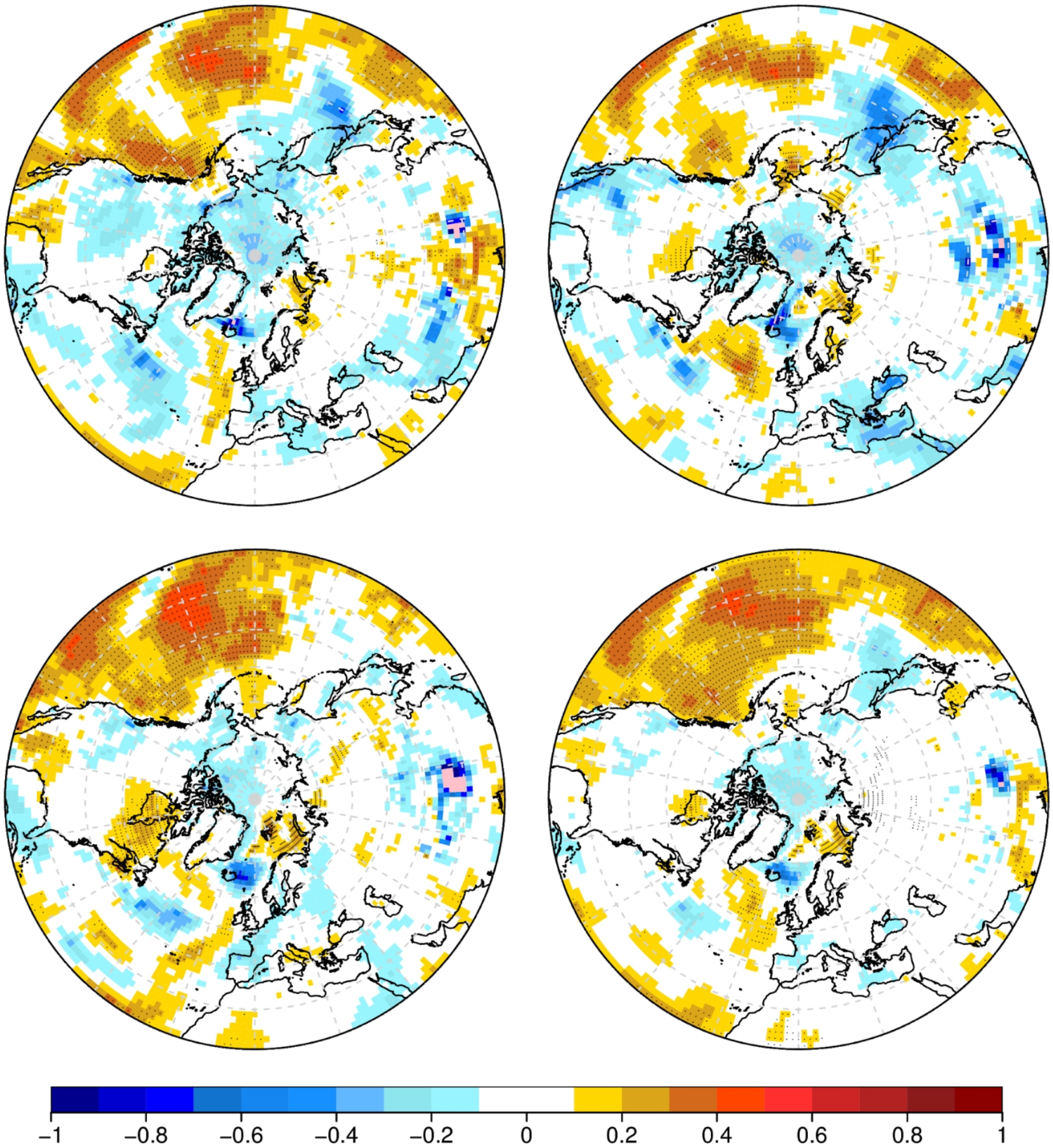


Fig 5.1.1: Fair continuous ranked probability skill score of the 1993-2014 DJF 2-meter temperature anomalies for ensemble simulations started in November 1st. Upper left (EC-Earth3.2 - 10 members), upper right (CNRM-CM6 - 10 members), lower left (GloSea5 C3S, 10 members), lower right (Multi-model). Dots indicate statistically significant values at a 95% confidence. Comparisons against ERA-Interim.

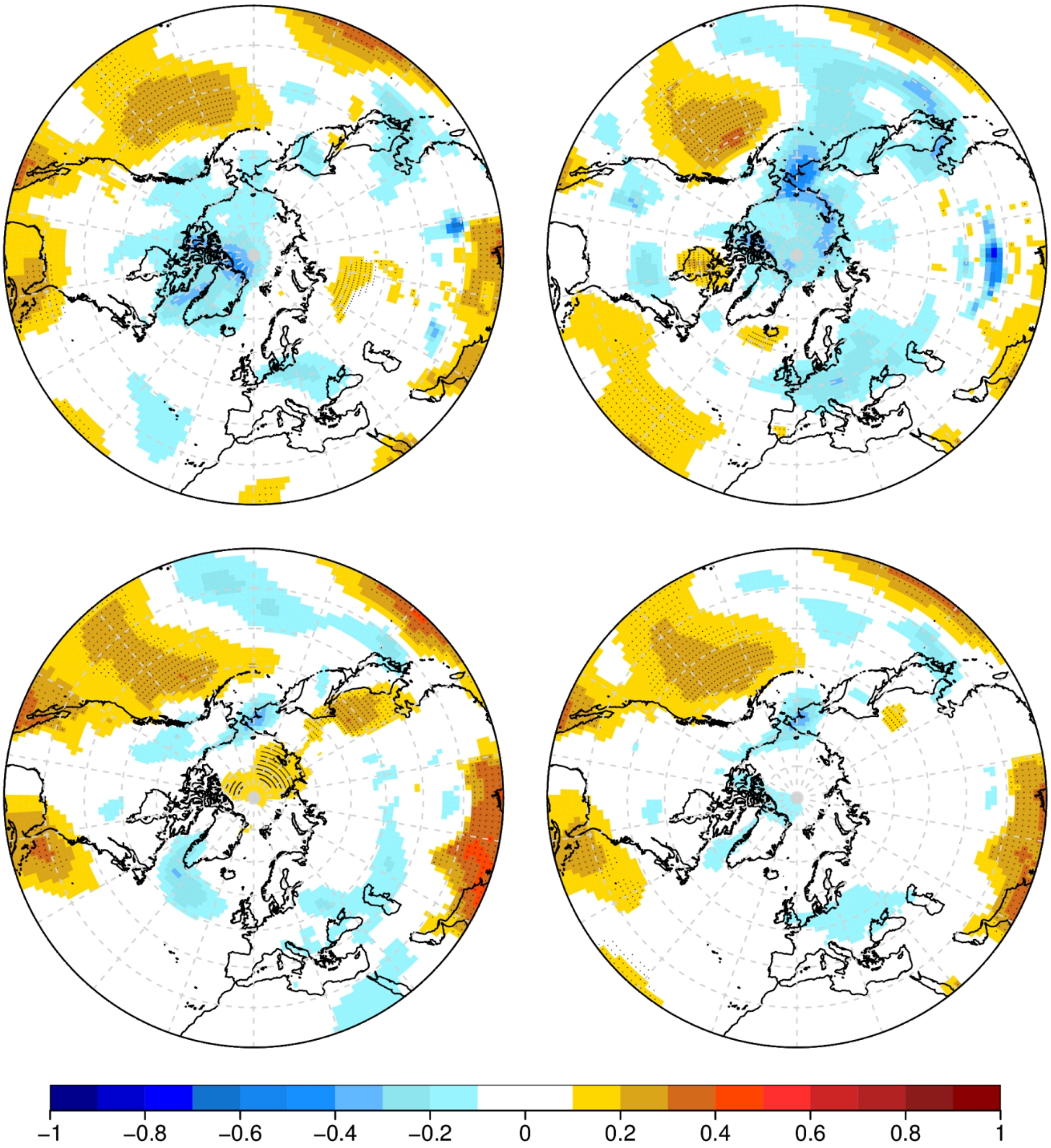


Fig 5.1.2: Fair continuous ranked probability skill score of the 1993-2014 DJF sea level pressure anomalies for ensemble simulations started in November 1st. Upper left (EC-Earth3.2 - 10 members), upper right (CNRM-CM6 - 10 members), lower left (GloSea5 C3S, 10 members), lower right (Multi-model). Dots indicate statistically significant values at a 95% confidence. Comparisons against ERA-Interim.

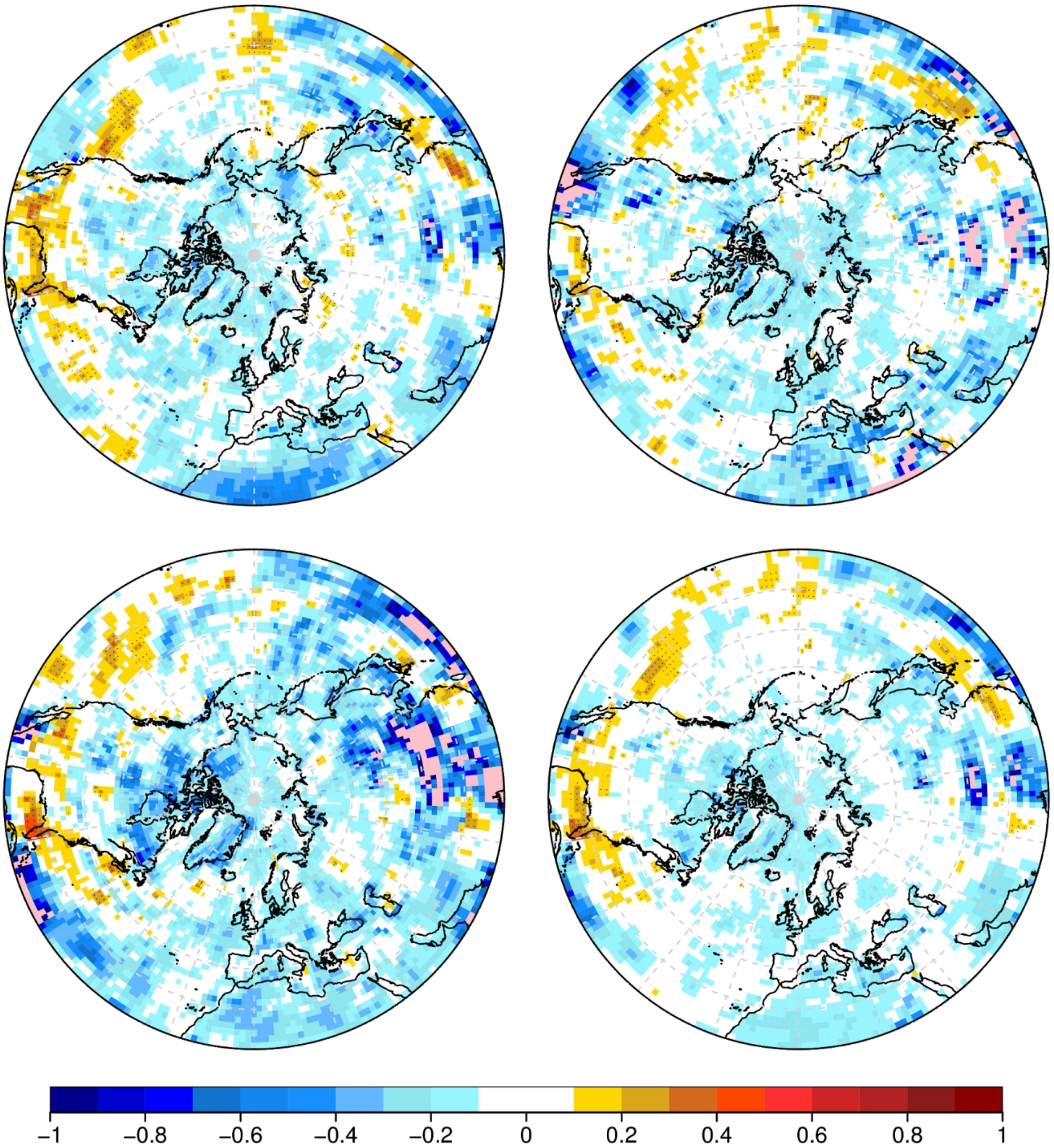


Fig 5.1.3 Fair continuous ranked probability skill score of the 1993-2014 DJF precipitation anomalies for ensemble simulations started in November 1st. Upper left (EC-Earth3.2 - 10 members), upper right (CNRM-CM6 - 10 members), lower left (GloSea5 C3S, 10 members), lower right (Multi-model). Dots indicate statistically significant values at a 95% confidence. Comparisons against GPCP V2.2.

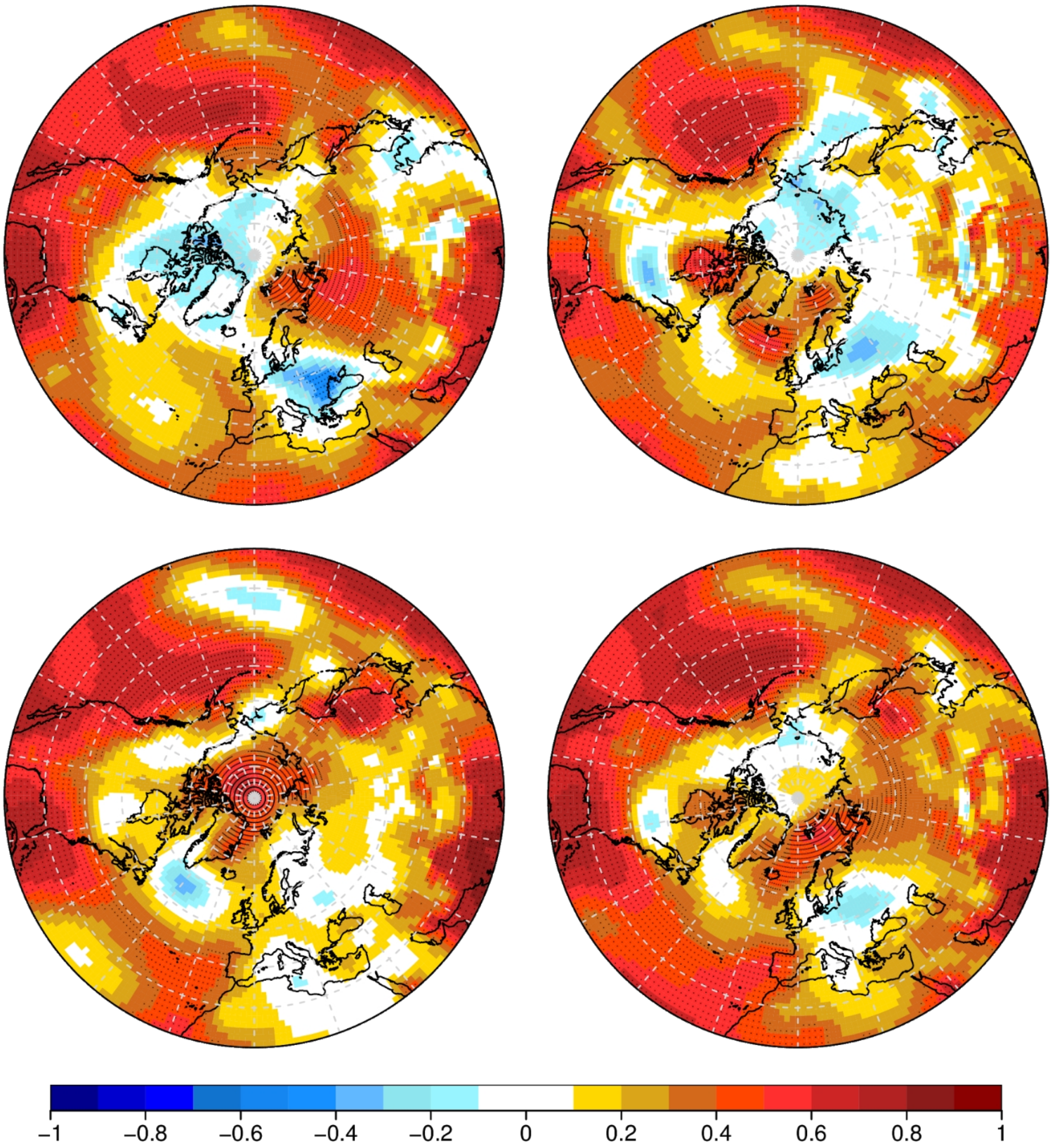


Fig 5.1.4 Anomaly correlation coefficient of the 1993-2014 DJF sea level pressure for ensemble simulations started in November 1st. Upper left (EC-Earth3.2 - 10 members), upper right (CNRM-CM6 - 10 members), lower left (GloSea5 C3S, 10 members), lower right (Multi-model). Dots indicate statistically significant values at a 95% confidence. Comparisons against ERA-Interim.

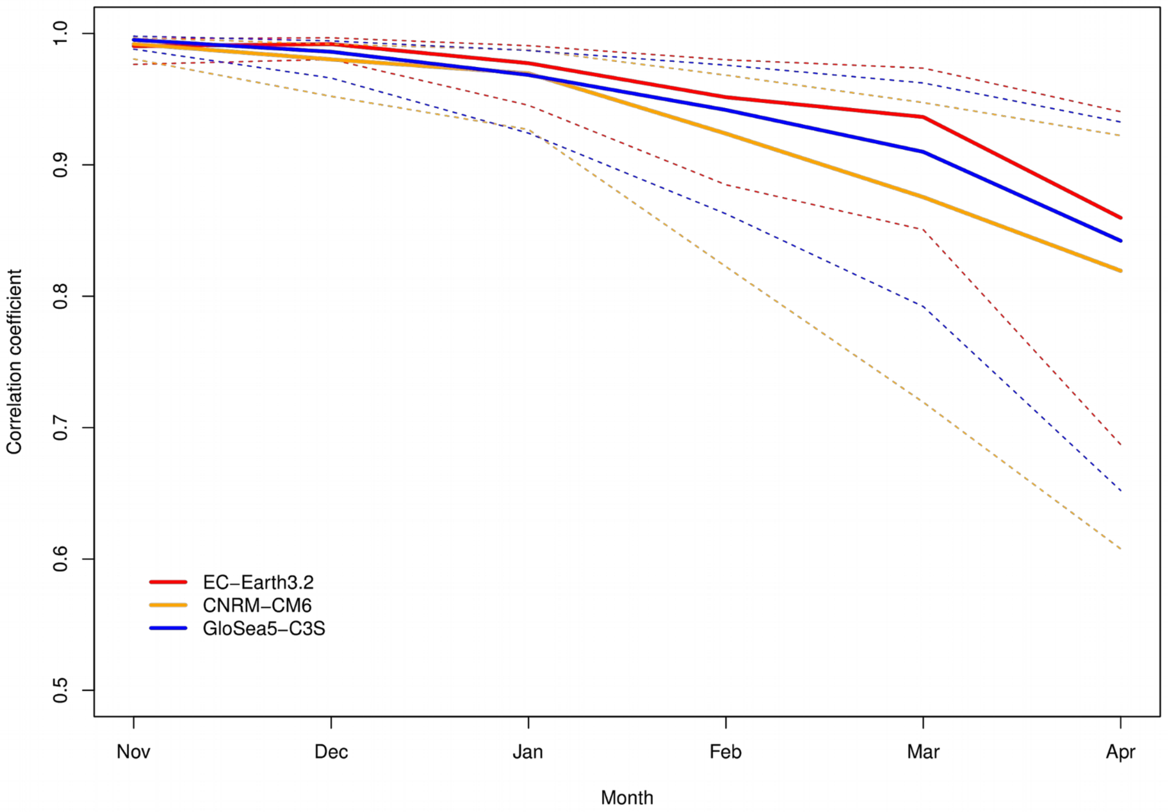


Fig 5.1.5: November 1st initialized 10-member ensemble forecasts of NINO3.4 index for the period 1993-2014. Comparison against ERA-Interim. Dashed lines show the 95% confidence intervals of the correlation coefficient. The confidence interval is computed by a Fisher transformation and the significance level relies on a one-sided student-T distribution.

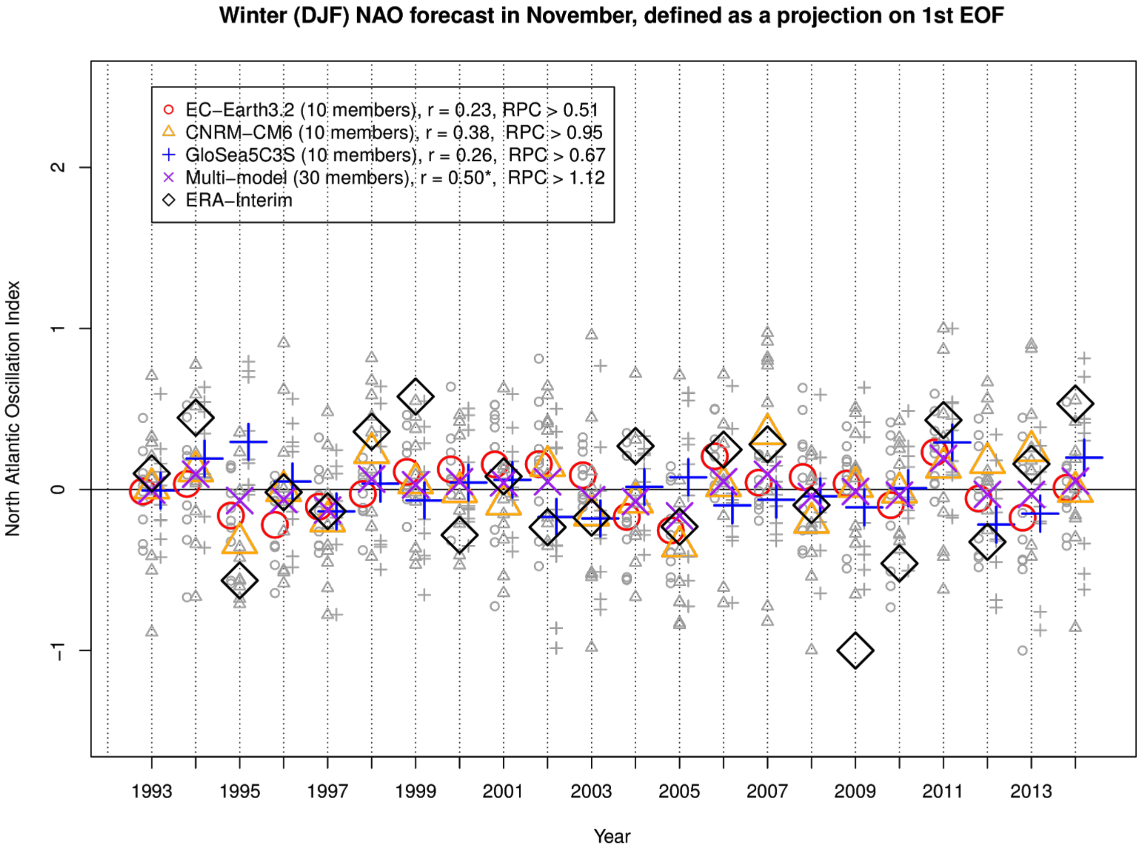


Fig 5.1.6: November 1st initialized 10-member ensemble forecasts of the normalized winter (DJF) North Atlantic Oscillation Index (NAOI) for the period 1993-2014, defined as a projection on 1st EOF of sea level pressure. Correlation coefficient values (r) versus ERA-Interim (Significance indicated by a star). Gray symbols display results from individual ensemble members. RPC stands for ratio of predictable components.

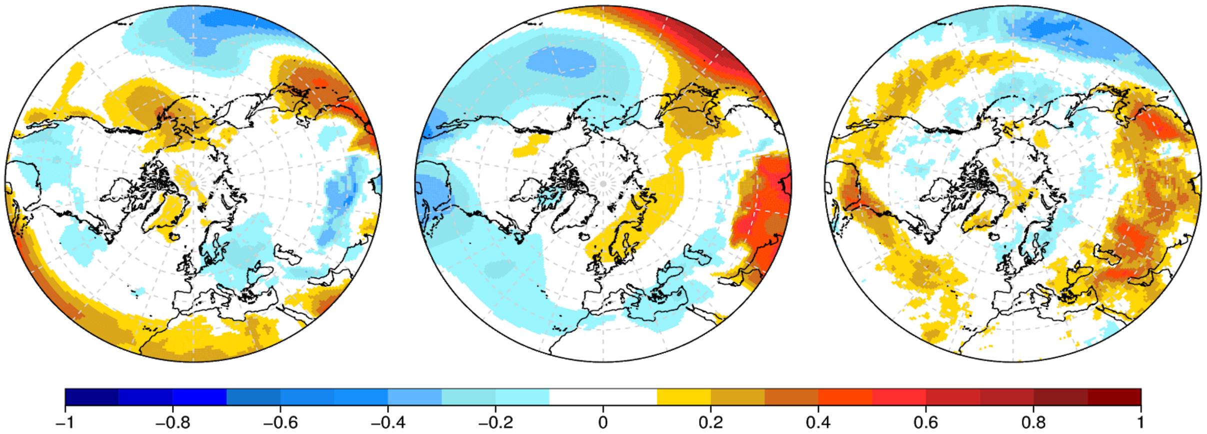


Fig 5.1.7: Surface temperature (left), sea level pressure (center) and precipitation (right) anomaly correlation with the NINO3.4 index for a 200 year control pre-industrial simulation with EC-Earth3.2. Statistically non-significant values are in white.

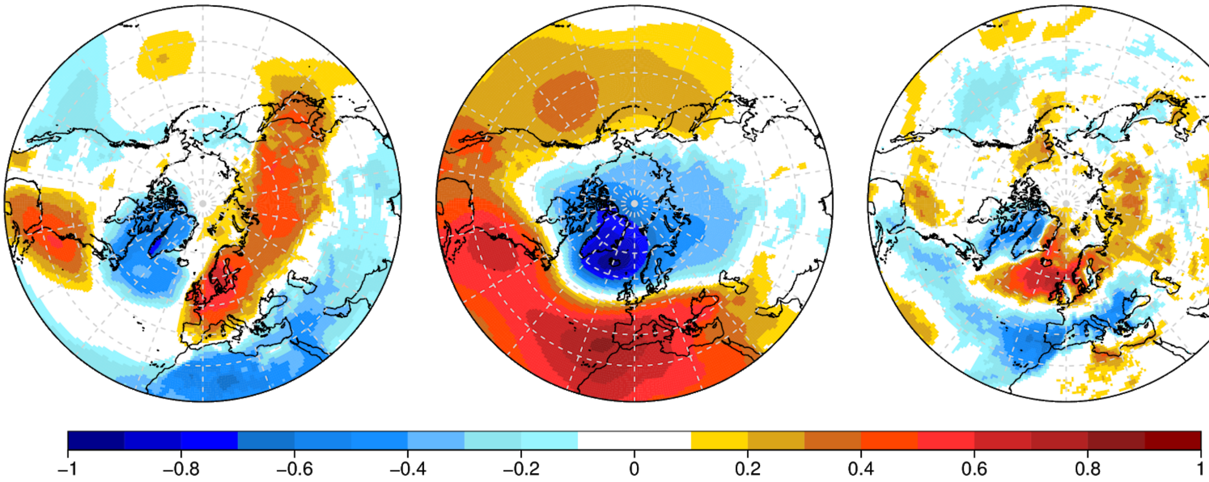


Fig 5.1.8: Surface temperature (left), sea level pressure (center) and precipitation (right) anomaly correlation with the station based NAO index for a 200 year control pre-industrial simulation with EC-Earth3.2. Statistically non-significant values are in white.

5.2. Skill of seasonal forecasting systems in representing summer Arctic sea ice

This section describes the ability and deficiencies of current state-of-the-art seasonal forecasting systems in reproducing Arctic sea ice concentration variability. The analysis is focused on coupled seasonal re-forecasts from the stream 1 of APPLICATE WP5 experiments (see Deliverable 5.1), as well as the ECMWF operational seasonal forecasting system 5 (SEAS5), for the summer season in May initializations. Some corresponding analyses for the winter season are provided in the annex to this deliverable.
Reference data for this analysis is NSIDC version 4.

5.2.1. Systematic errors in sea ice concentration and extent

A preliminary step in evaluating the forecast quality of these models is to compute systematic errors in the raw model outputs for sea ice concentration.

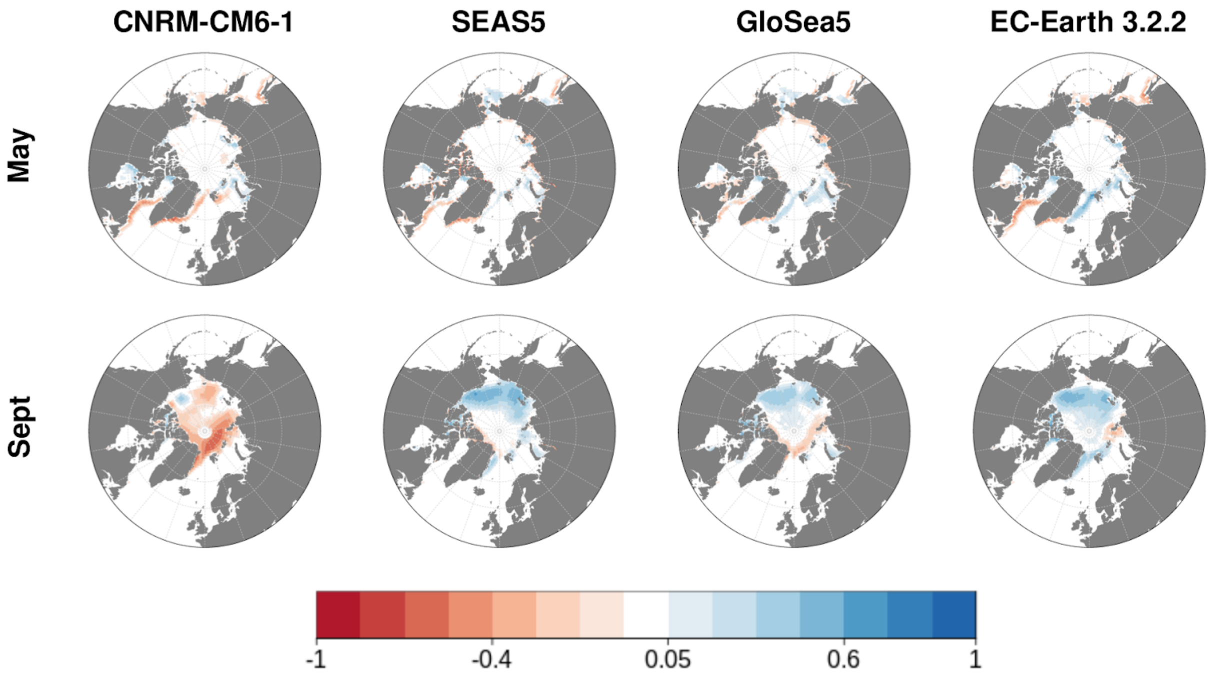


Fig. 5.2.1: Mean bias in monthly mean sea ice concentration with NSIDC v4 in May (forecast month 1) and September (forecast month 5) for each of the coupled systems.

Fig. 5.2.1 shows the mean bias over the re-forecast period of month 1 (May) and month 5 (September) SIC with respect to NSIDC. Red areas show where SIC is too low in the models, whereas blue areas highlight where model have excessive SIC. From the first month of simulation, the systems exhibit different behaviors. CNRM-CM6-1 has too low SIC along the ice edge in the Labrador and Greenland seas, whereas the other systems show too high SIC in the Iceland and Nordic seas. At longer lead times, CNRM-CM6-1 exhibits a substantially different bias than the other models, with too little SIC over most of the Arctic. This is due to the initialization strategy for this system, for which even at the initial stage, sea ice thickness is often too low. During the melt season, this results in an excessive reduction of SIC over most of the Arctic, and a subsequent loss in predictability (see later evaluations).

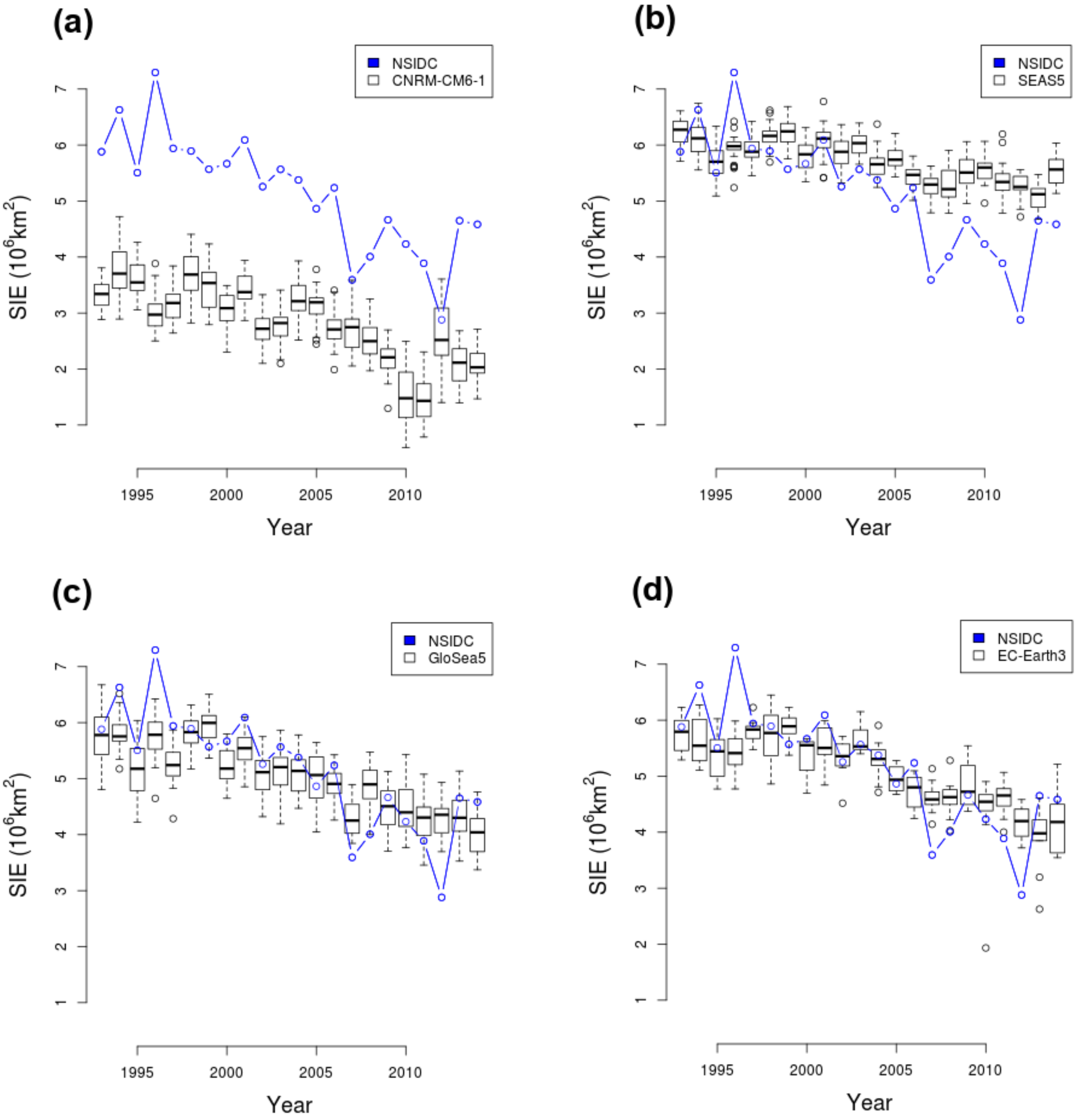


Fig. 5.2.2: Boxplots representing SIE values for the re-forecast ensembles in each of the models (a) CNRM-CM6-1, (b) SEAS5, (c) GloSea5 and (d) EC-Earth3 compared to NSIDC data (in blue). Boxes show the inter-quartile range of the ensembles, the thick black line is the ensemble median, whiskers show the range of the ensemble up to 1.5 σ, and dots represent outliers beyond this range.

Figure 5.2.2 shows for each year of the re-forecast period the interquartile range and spread of ensemble members (non bias-corrected model outputs) in May re-forecasts for September SIE, compared to NSIDC reference data. The systems exhibit different characteristics: while CNRM-CM6-1 clearly underestimates SIE for most years of the re-forecast, but seems to capture the negative trend in SIE over 1993-2014, SEAS5 shows values comparable to NSIDC in the beginning of the re-forecast period but tends to underestimate the negative trend, leading to an overestimation of September SIE for all years after 2006. Both GloSea5 and EC-Earth3 remarkably capture the overall negative trend, and NSIDC values are inside the range of the ensemble for most years of the re-forecast in these two systems.

From this analysis, it appears crucial to bias-correct the SIC values for these systems, and remove the trend in the consecutive skill evaluations so as to avoid overestimating actual skill of the models.

5.2.2. Pan-Arctic sea ice extent

To get a first glimpse of the skill of different systems in re-forecasting sea ice conditions, we focus on Pan-Arctic sea ice extent (SIE) computed with a 0.15 sea ice concentration threshold. RMSE and correlation over the 1993-2014 re-forecast period are shown in Fig. 5.2.3. These skill scores are calculated for linearly-detrended sea ice concentration data so as to avoid over-estimating the skill due to the strong negative trend in sea ice extent.

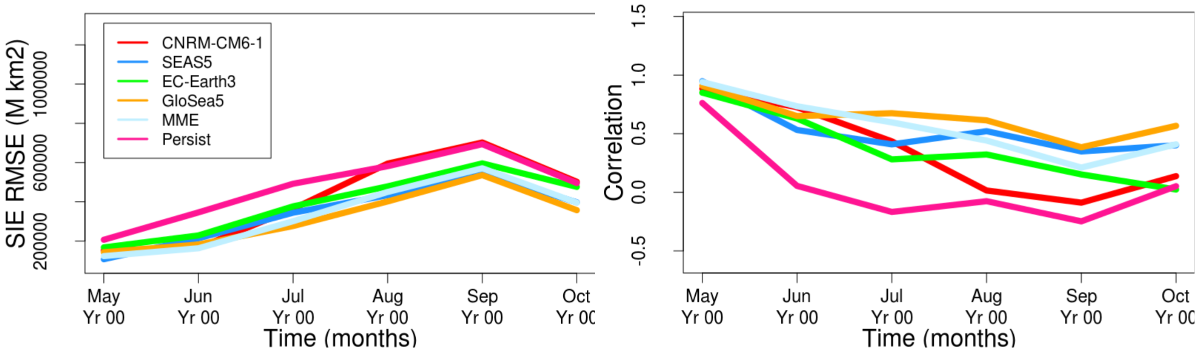


Fig. 5.2.3: Evolution according to forecast month of pan-Arctic SIE RMSE (left) and correlation (right) with NSIDC reference data. The multi-model ensemble (MME) is shown in light blue, and persistence of April anomalies in pink.

The skill of individual systems is compared to a multi-model ensemble (MME) grouping all ensemble members of each system together (without weighting individual systems but with equal weight for each member). The skill of the MME is shown in blue. Scores can be compared to a simple persistence approach (persisting SIC anomalies from April to the following months) for which results are shown in pink. Most systems exhibit fairly similar levels of skill, both for RMSE and correlation. RMSE is maximum in September when SIE is at the minimum of the seasonal cycle. Correlation drops (as expected) with lead times from over 0.8 in May to near-zero correlation for two of the models in October, namely CNRM-CM6-1 and EC-Earth 3.2. The other two systems, namely SEAS5 and GloSea5, still exhibit significant levels of correlation with NSIDC data at a 6-month lead time. All models show higher skill than persistence, although the score for persistence is inside the range of uncertainty after 2 months lead time in most cases, likely due to the limited number of re-forecast years in the evaluation (not shown).

5.2.3. Sea ice edge forecast quality

While seasonal forecasts of Pan-Arctic sea ice can provide some indication of below-average or above-average presence of sea ice, these are not the most relevant indicators for potential end-users of seasonal forecast information. Among these users, some are most interested in probabilities of presence of sea ice along shipping routes or near the climatological sea ice edge (Melia et al. 2017).

We therefore evaluate the skill of the different models in representing the position of the sea ice edge (based on monthly averages) by computing the IIEE metric introduced by Goessling et al. (2016). This is done after correcting SIC for systematic errors with a simple cross-validation bias removal. This simple method has some caveats, since for bounded fields such as SIC values it can yield values outside the theoretical range. More elaborate methods exist such as that of Dirkson et al. (2018), but are beyond the scope of this deliverable.

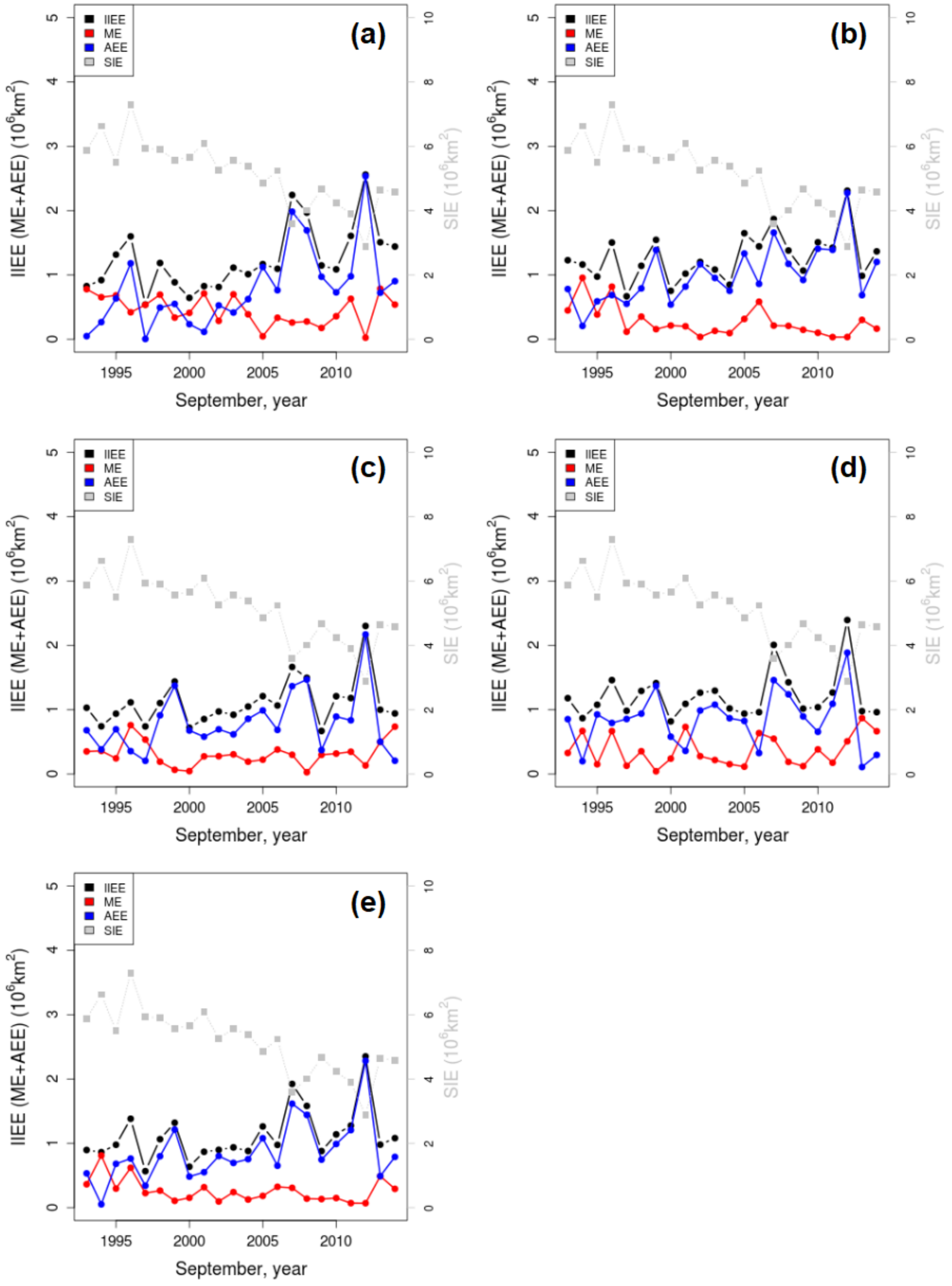


Fig 5.2.4: IIEE (black, in millions of km2) and decomposition in ME (red) and AEE (blue) with respect to NSIDC data for September 1993 to 2014 in re-forecasts initialized in May with (a) CNRM-CM6-1, (b) SEAS5, (c) GloSea5 and (d) EC-Earth3. (e) Same as (a-d) but for a multi-

model ensemble grouping all ensemble members of each individual system (after individual bias correction of SIC). The grey line shows the reference SIE (y-axis on the right hand side).

Figure 5.2.4 shows the IIEE for each individual system for September 1993-2014, as well as for a multi-model ensemble grouping each individual member of each system (after bias correction) into a large ensemble. Results for the different systems are quite similar, with IIEE increasing during the re-forecast period, mainly due to an increase in AEE.

Peaks in IIEE are found in 2007 and 2012 for each system. Some systems, namely CNRM-CM6-1 and EC-Earth 3, show more variability in the misplacement error than SEAS5 and GloSea5. This suggests that for the latter, skill evaluations based on RMSE of Pan-Arctic SIE are giving a rather correct picture of the model capacity to predict the sea ice edge position, whereas for the former two systems, the Pan-Arctic SIE actually “hides” some compensation between areas where SIC is overestimated and where it is underestimated.

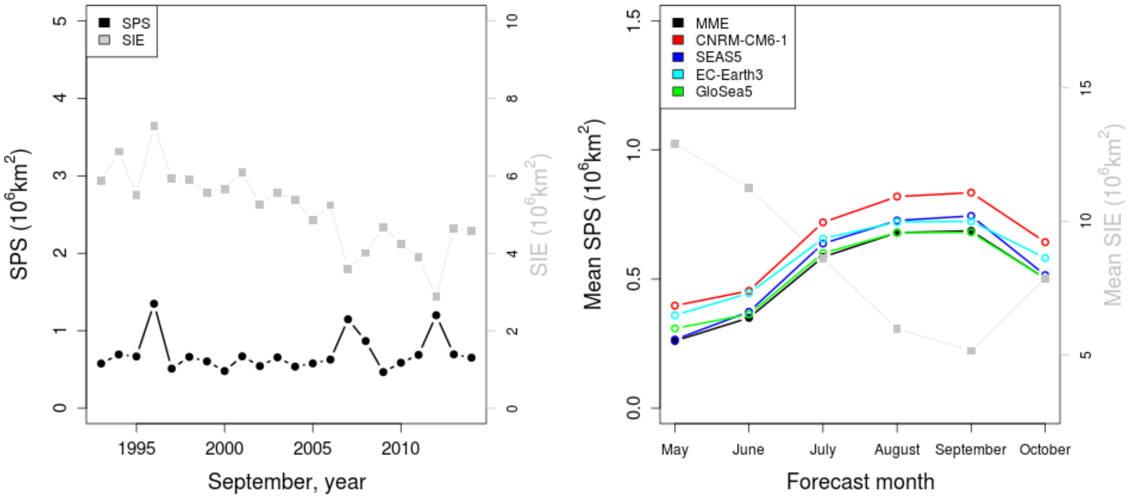


Fig. 5.2.5: Right: SPS for September 1993-2014 over the Pan-Arctic region in the MME of re-forecasts initialized in May (in black), and total SIE (in grey, right y-axis); left: mean SPS over 1993-2014 according to forecast month for each system and the MME (in black).

Figure 5.2.5 shows the SPS for monthly SIC in September over the re-forecast period for the multi-model ensemble. Little inter-annual variability in SPS is found, setting aside the 1996, 2007 and 2012 cases during which SIE over the Pan-Arctic region reached local extrema (maximum in the case of 1996, minima in 2007 and 2012, as shown in grey in the figure). The right hand side figure shows the average SPS over the 1993-2014 period for each re-forecast month for each system (colors) and the multi-model ensemble (black). Unlike previous results for atmospheric fields over other regions (see e.g. Hagedorn et al. 2005), the multi-model approach does not significantly improve results with respect to the best single models. Each system exhibits quite similar mean values over the re-forecast period, and similar peaks to those of the MME are found for each system in September (not shown). The CNRM-CM6 model exhibits once more some difficulties related to the too thin sea ice initialization, leading to a loss of predictability in August to October.

5.2.4. Illustration of model deficiencies using a case study: September 2012

This section focuses on the seasonal re-forecasts for September 2012 sea ice extent, which corresponds to the minimum over the 1993-2014 period for Pan-Arctic sea ice extent. This is to provide an illustration of the model deficiencies and discuss the limitations of the simple bias correction approach used in this deliverable. Figure 5.2.6 shows the sea ice edge for September 2012 in the ensemble mean of each individual system initialized in May, defined using a SIC > 0.15 threshold. This is compared to the ice edge obtained with NSIDC data. As in Figure 1 in Goessling et al. (2016), we depict in blue areas where SIC is overestimated beyond this 0.15 level, leading to a too extended sea ice edge, and in red areas where SIC is underestimated leading to a too restricted sea ice edge contour. Some of the systematic errors in September illustrated in Fig. 5.2.1 can be found for this specific forecast date: CNRM-CM6-1 has too low SIC in the Central Arctic, leading to the underestimation of the ice edge in the Greenland seas and close to the pole. SEAS5, GloSea5 and EC-Earth 3.2 exhibit a typical extension of SIC too far towards Alaska and Eastern Siberia, therefore overestimating the sea ice edge in the Beaufort, Chukchi and East Siberian seas.
The Atlantic sector is generally better forecasted in the different systems than in the Pacific.

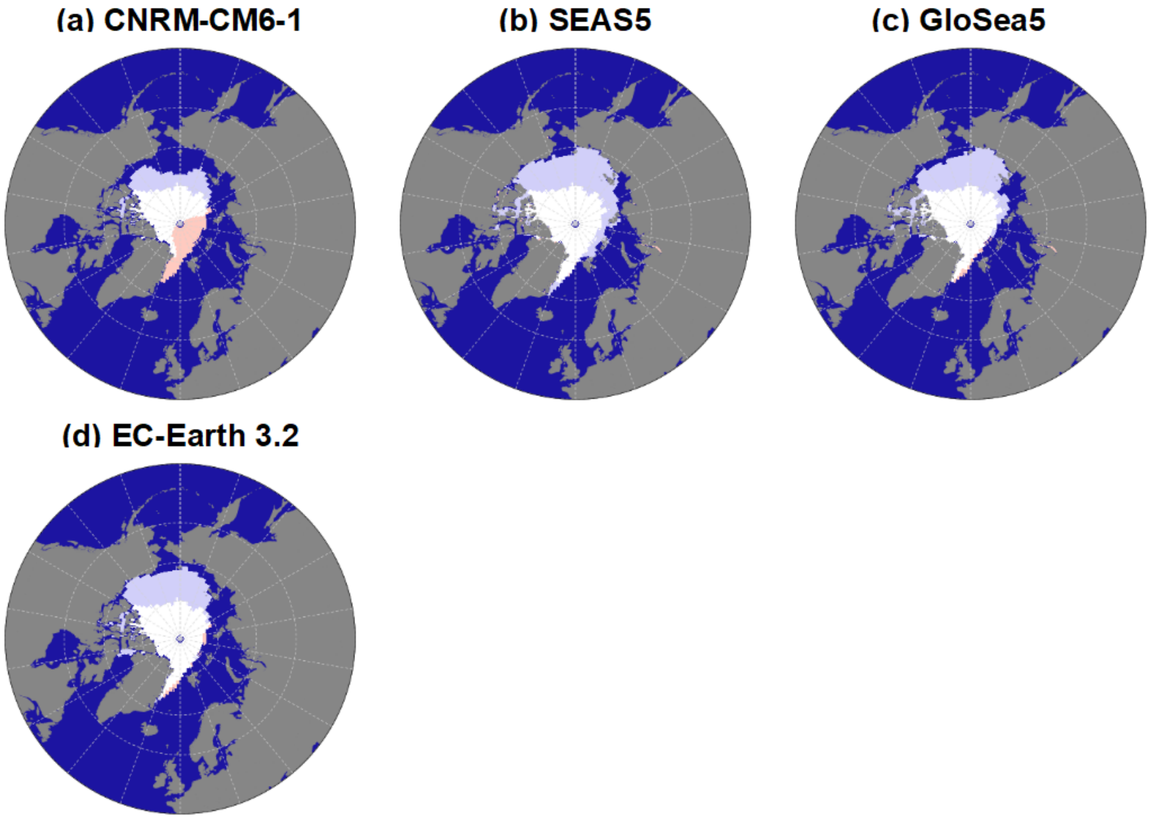


Fig. 5.2.6: Raw Arctic sea ice edge re-forecasts for September 2012 in each system, compared to NSIDC. Areas in white indicate where sea ice is found in both NSIDC and the ensemble mean re-forecast. Areas in red show where sea ice was found in NSIDC but wasn't forecast by the model, while areas in light blue highlight regions where the sea ice was forecast in the model but not present in NSIDC. The threshold for SIC used here is 0.15.

These raw forecasts can be compared to those obtained after a straightforward bias correction of SIC fields using leave-one-out cross-validation. These are shown in figure 5.2.7. Despite very different biases between CNRM-CM6-1 and the other systems, all models after bias correction over-estimated the total pan-Arctic sea ice extent, mainly due to a clear extension of the sea ice edge towards the Bering Strait.

When looking at the forecasts for September 2012 in figure 5.2.2 and their corresponding ensemble spread, compared to that of surrounding years, it appears that no individual system clearly singled out 2012 as a year of record-low SIE.

This particular year appears therefore as a clear "forecast bust", and reasons for such misses in the seasonal forecast systems despite fair correlation levels for total SIE should be further investigated in the framework of the project. In the case of 2012, part of the decline in SIE was however attributed to a strong storm in August (Parkinson and Comesino (2013)), which limits the extent to which such a minimum can be predicted at extended time scales. Events at a sub-seasonal time scale may therefore significantly alter the quality of seasonal predictions if these are not at least partially captured by the ensemble.

However, beyond the actual predictability of the 2012 minimum, this evaluation does highlight the need for significant improvements in the systematic errors and drift in the sea ice component of the coupled models evaluated in this deliverable.

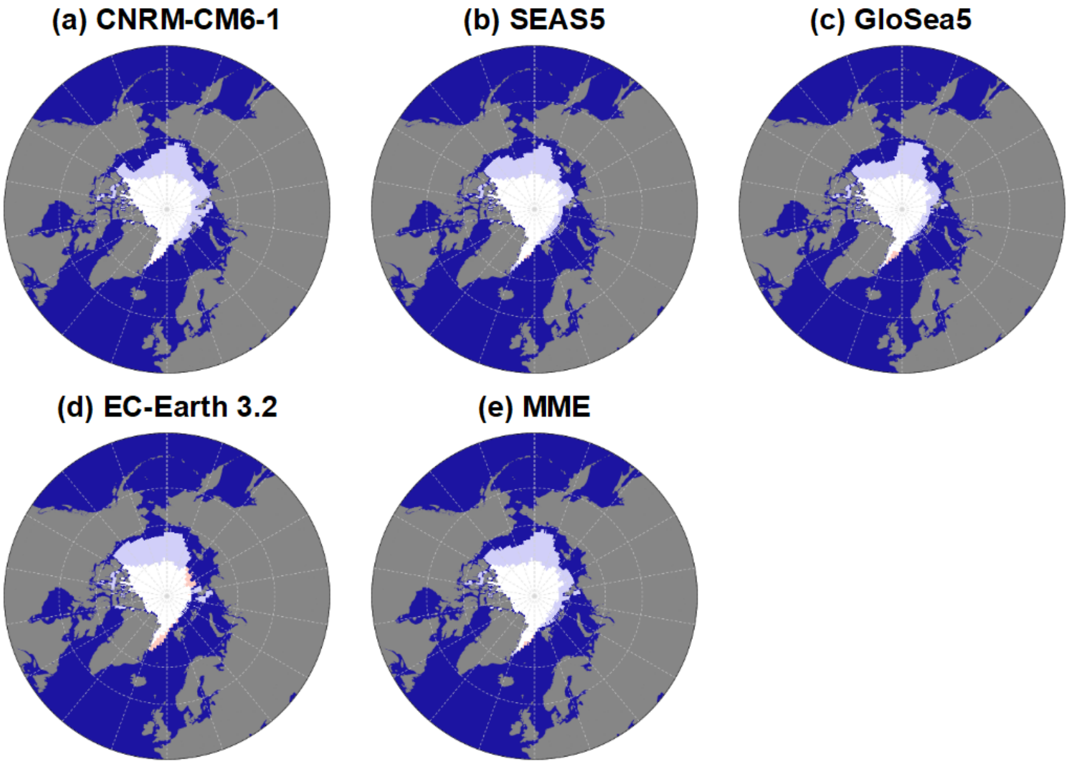


Fig. 5.2.7: Same as fig. 5.2.6, but for bias-corrected September 2012 sea ice edge over the Arctic in each re-forecast compared to NSIDC.

5.3. Statistical forecasting

Section 5.3 evaluates the skill of statistical empirical models for predicting Arctic sea ice area (SIA) and sea ice volume (SIV) anomalies as proposed in Task 5.2.4. This ongoing task is mainly guided by three well-defined goals, as follows: (i) Assess the skill of multiple linear regression models for predicting the Arctic sea ice area and volume anomalies at two critical months of the year (March and September), for lead periods of up to twelve months; (ii) Determine whether, and eventually how, model resolution can play a role on the statistical predictability of the Arctic sea ice? (iii) Identify whether or not the statistical predictability of sea ice area and volume anomalies is losing skill over time.

To do so, we make use of four different GCMs with two configurations each (“low” and “high” resolutions), totalizing eight model simulations. Namely, the models are: HadGEM (Roberts et al., 2018a; Williams et al., 2018), ECMWF-IFS (Roberts et al., 2018b), AWI-CM (Wang et al., 2014; Sidorenko et al., 2015) and MPI-ESM (Müller et al., 2018). Model outputs come from HighResMIP (Haarsma et al, 2016), which is one of the CMIP6-endorsed Model Intercomparison Projects (MIPs). These results were previously used in the context of PRIMAVERA (Process-based climate sIMulation: AdVances in high-resolution modelling and European climate Risk Assessment) project (e.g., Docquier et al., 2018), which is another European Union Horizon2020 Project. This task is being developed in close collaboration with PRIMAVERA and it is a successful example of clustering between two Horizon2020 projects.

The empirical models are built mainly based on recommendations from Drobot et al. (2006) and Lindsay et al. (2008). They are generated following a Monte Carlo (MC) method, with 500 realizations at every step. For each realization, 80% of the data are randomly selected and used to build the statistical model, while the remaining 20% of the data are used for verification, estimating the errors (RMSE) from statistically reconstructed values and model outputs. Fig. 5.3.1 displays the first MC realization used to construct three different statistical models for predicting the SIV in September (SIV_{SEP}), having as predictor the SIV_{JAN} and SIA_{JAN} in January (8-month lag), as well as the combination of these two variables. Fig. 5.3.2 shows the RMSE estimated at each MC realization, for all lagged-months.

For the example mentioned above, with SIV_{SEP} predicted by SIV itself and SIA, Fig. 5.3.3 reveals that SIA does not substantially improve the skill of the statistical model and the SIV_{SEP} is more efficiently predicted by its own values at lagged-months. This observation is valid for the eight model configurations.

Fig. 5.3.4 compares the RMSE resulting from the statistical model performed with two different configurations of the same GCM. Interestingly, the skill of the statistical models is better for the high-resolution version, except for MPI-ESM in which the statistical skill is slightly better for the low-resolution version. However, such better skill for high-resolution models tend to attenuate near September (also observed in Fig. 5.3.2). Finally, the results also suggest that the statistical predictability substantially improves in July, at about 3-month lag. A feasible explanation for these results has not yet been found and would require further investigation.

In total, seven predictors are being considered for building the multiple linear regression models: Sea Ice Area (SIA), Sea Ice Volume (SIV), Sea Ice Concentration (SIC), Sea Ice Thickness (SST), Sea Surface Temperature (SST), Ice Velocity (Drift) and the poleward Ocean Heat Transport (OHT). To the knowledge of the authors, this is the first time that poleward OHT is included as predictor for sea ice parameters. At the moment of closing this report, results are being analyzed and improved, in order to be incorporated to a scientific manuscript in the upcoming months.

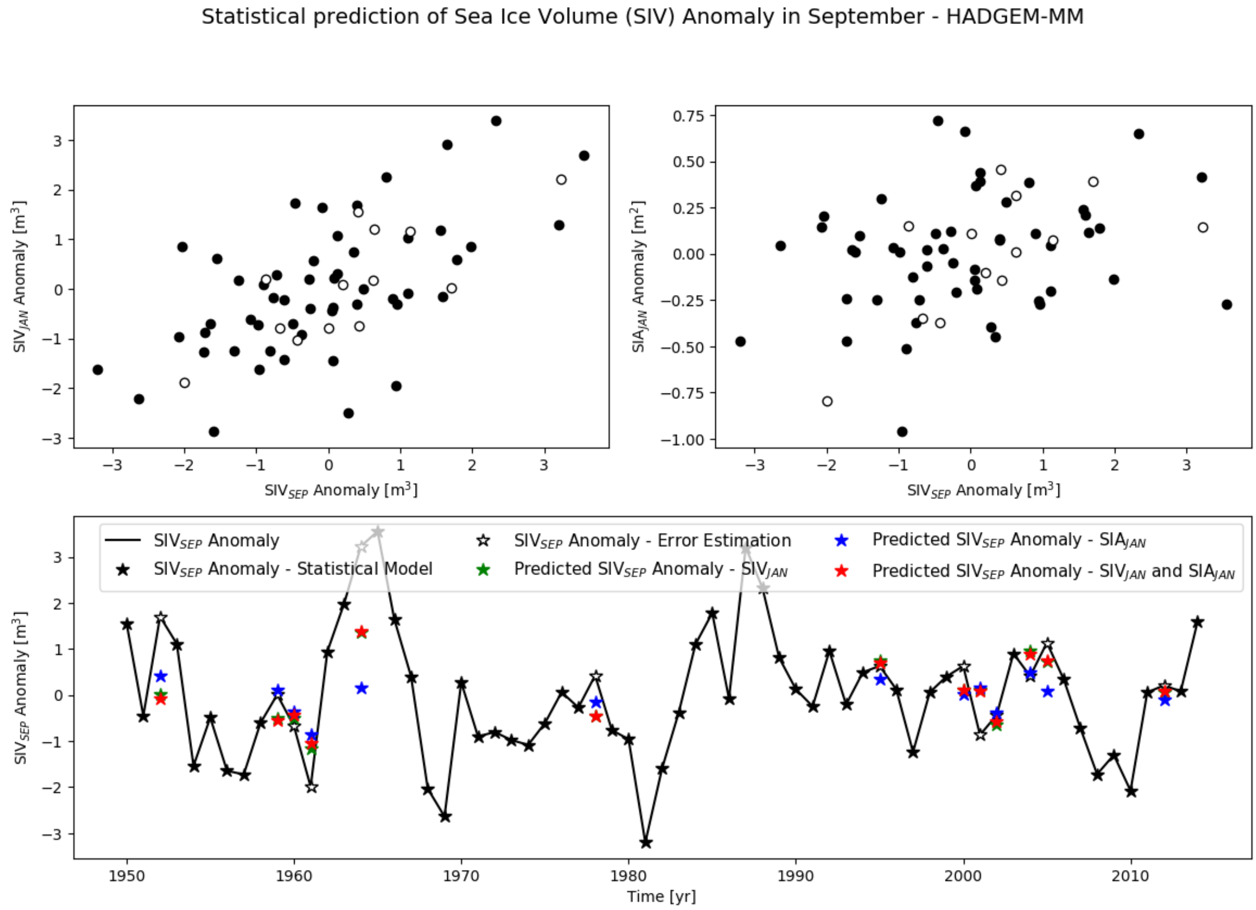


Fig. 5.3.1: (Top) Diagrams SIV_{SEP} x SIV_{JAN} (left) and SIV_{SEP} x SIA_{JAN} (right). Black circles represent the data used to built the statistical model, while white points indicate the data used for calculating the RMSE and testing the skill of the statistical model. (Bottom) Black line represents the SIV_{SEP} time series. Black and white stars are equivalent to the black and white circles described above, respectively. Green, blue and red stars display the statistically predicted values for SIV_{SEP} having as predictor SIV_{JAN}, SIA_{JAN} and SIV_{JAN}+SIA_{JAN}, respectively.

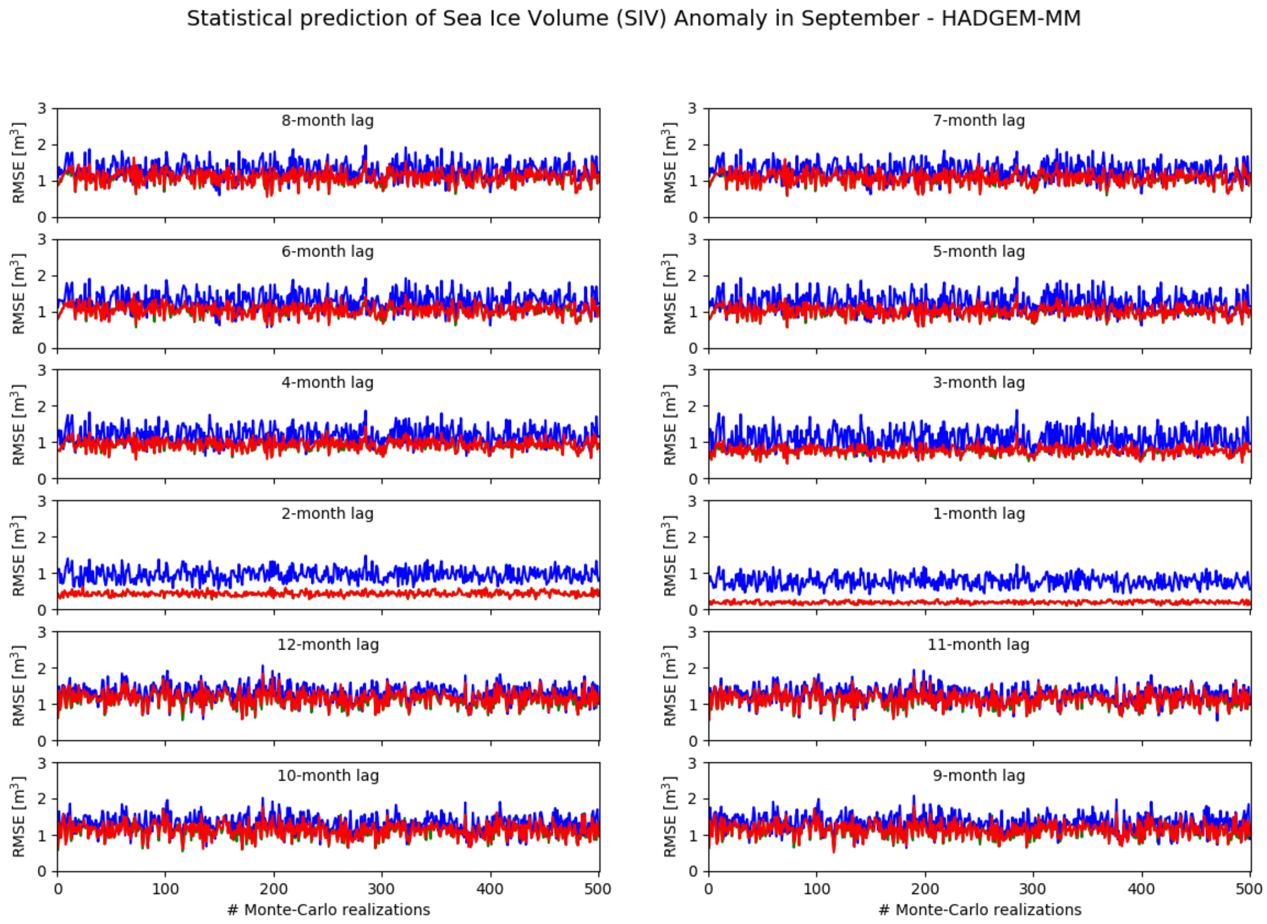


Fig. 5.3.2: Root Mean Squared Error (RMSE) estimated between the statistically predicted values and the respective model outputs selected for testing the statistical model at every Monte Carlo realization. Blue, green (overlapped by the red) and red represent the RMSEs from the models which have as predictor SIV_{MON} , SIA_{MON} and SIV_{MON}+SIA_{MON}, respectively. The subscript index "MON" refers to the predictor month. For instance, 8-month lag is January, 7-month lag is February, etc. The 12 to 9 month-lags predict the SIV in September of the next year.

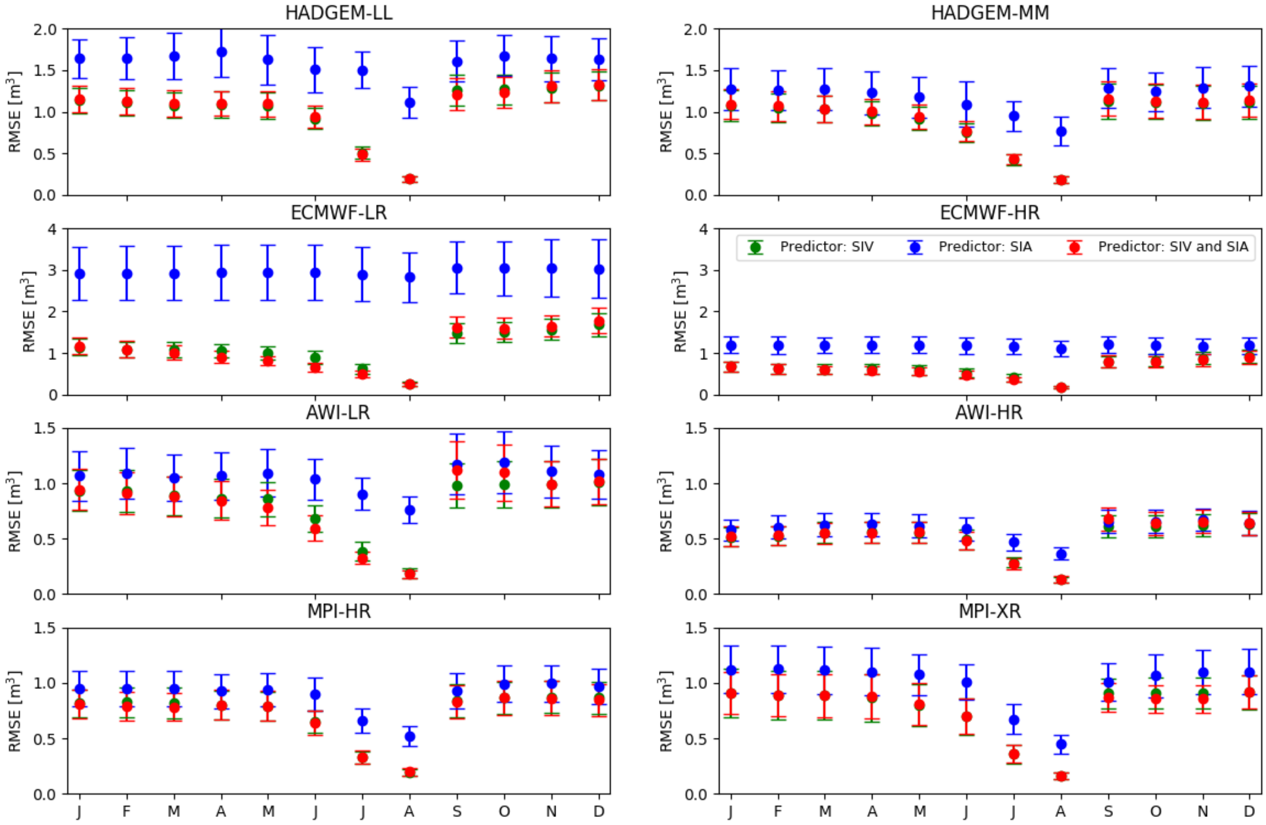


Fig. 5.3.3: Monthly mean Root Mean Squared Error (RMSE) estimated from all Monte Carlo realizations. The error bar represents the standard deviation. Blue, green and red represent the RMSEs from the models which have as predictor SIV_{JAN} , SIA_{JAN} and SIV_{JAN}+SIA_{JAN}, respectively. Left panels represent the "low" resolution versions of the models, while right panels represent the high resolution configurations. Notice that the MPI-HR is our "low" resolution version of MPI-ESM model.

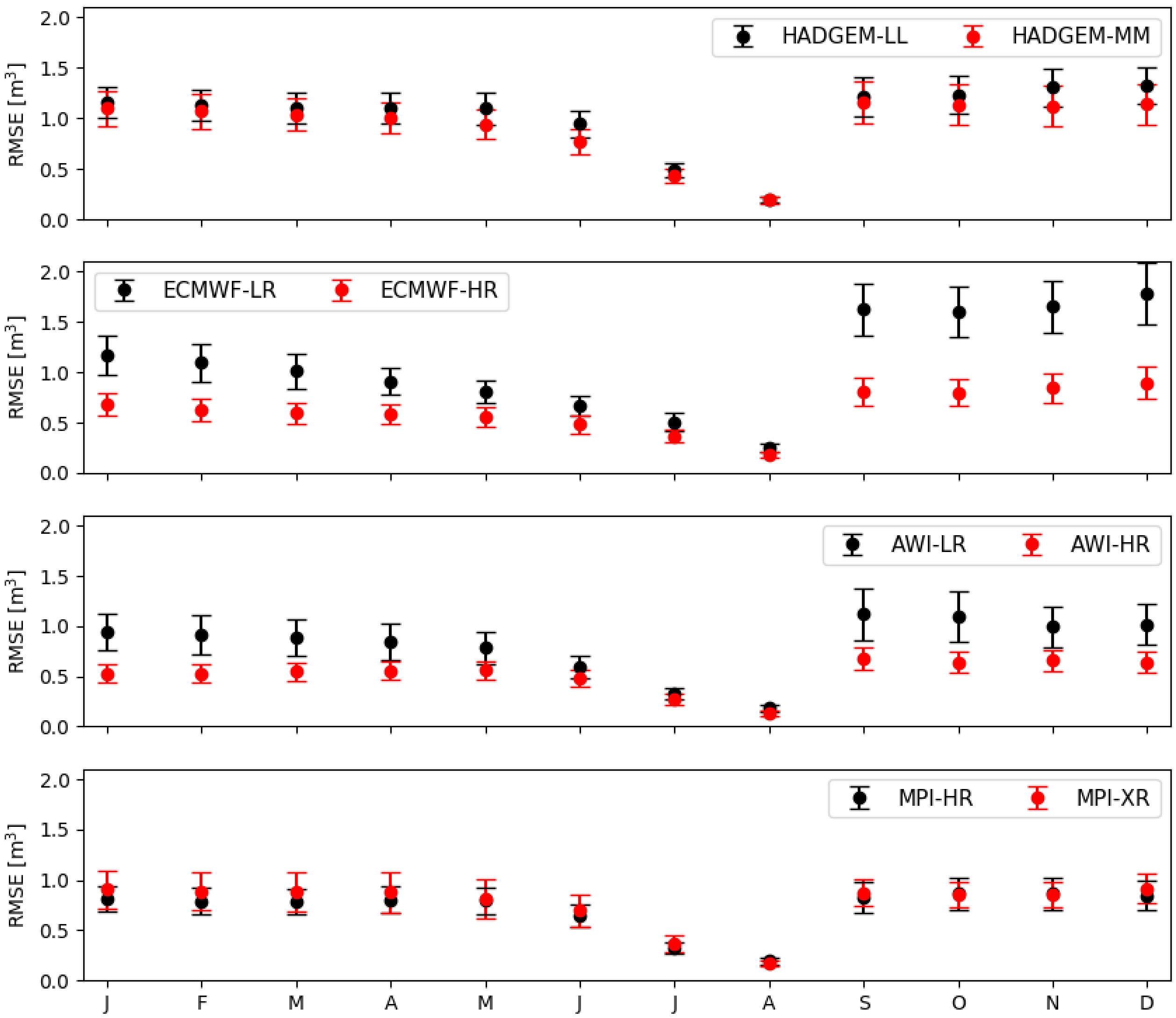


Fig. 5.3.4: Monthly mean Root Mean Squared Error (RMSE) estimated from all Monte Carlo realizations. The error bar represents the standard deviation. Black and red represent the “low” and “high” resolution configuration of the models, respectively.

6. CONCLUSIONS AND OUTLOOK

In this report, we have shown a comprehensive evaluation of forecast quality over the Arctic and mid-latitudes, based on a set of state-of-the-art models focusing on weather to seasonal climate prediction time scales.

The first part of this deliverable focused on a case study for extreme precipitation in Svalbard which occurred from 7-10 November 2016, and more generally on several events during the recent 2016-2018 period. A comparison between the AROME-Arctic limited area model and ECMWF deterministic global HRES forecasts highlighted the importance of a detailed orography to better capture precipitation amounts. However, the added value of high resolution is clearer when all days are included in the analysis than when subsetting the scores on high precipitation events only. AROME-Arctic was also found to have an improved ability to separate rain and snow events over the region.

The analysis was then extended to investigate added value in short range forecasts of high resolution limited area models compared to global models for a region encompassing all Norwegian SYNOP stations for the March 2016 to April 2018 period. The added value of a high resolution model was found to depend on the variable, type of station (e.g. inland or mountain) and the season of verification.

At the medium range, ECMWF forecasts are found to be on average less skillful over the Arctic than over the Northern Hemisphere (20°N-90°N), although the rate of improvement of 500 hPa geopotential height correlation over the past two decades is very similar for both regions. The evaluation of 2016-2017 winter and summer seasons medium range forecasts shows that errors for geopotential height in the troposphere are overall higher in summer than in winter. The spatial distribution of errors also varies with higher RMSE over the pole in summer, and a maximum over the north Atlantic in winter.

At the seasonal time scale, models from APPLICATE stream 1 as well as SEAS5 were evaluated and found to show a reasonable climatology in blocking frequency when compared to ERA-Interim, although this is no guarantee of actual skill in predicting a higher-than-normal blocking activity at a seasonal time scale. The evaluation of winter seasonal re-forecasts based on a fair continuous ranked probability skill score demonstrated the very limited ability of models to properly represent the variability of atmospheric fields such as sea-level pressure, temperature and precipitation at this time scale, with most skill arising from the predictability of ENSO. However, gridpoint correlation with ERA-Interim suggests some signal can be extracted from the ensemble mean of these forecasts when using larger ensembles and a multi-model approach.

Sea ice concentration forecast quality and skill was also examined in the seasonal re-forecasts. Beyond typical Pan-Arctic sea ice extent skill scores, we used deterministic and probabilistic integrated scores accounting for both total extent errors as well as misplacement errors to get a more complete overview of strengths and weaknesses of current systems. Levels of skill are quite similar between systems for both summer and winter seasons, with some sensitivity to the initialization strategy. Statistical forecasts based on numerical model experiments suggest that longer lead times for sea ice predictability than those found with current forecasting systems could be achieved, by improving initialization as well as moving to higher resolution.

Work is currently underway to improve both weather and seasonal climate prediction systems, by enhancing sea ice models and improving air-sea interactions (task 5.3.1), working on increased resolution (tasks 5.3.2 and 5.3.3), and improving ensemble generation (5.3.4).
Conclusions from these different areas of model improvement will be summarized in an upcoming deliverable (5.3), and the most promising approaches will be implemented in a second stream of experiments in the framework of the project to be compared with forecasts analysed here.

7. REFERENCES

Peer-reviewed literature

Adler, R.F., G. J. Huffman, A. Chang, R. Ferraro, P. Xie et al. (2003) The Version 2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979-Present). *J. Hydrometeor.*, 4, 1147-1167.

Athanasiadis, P.J., Bellucci, A., Scaife, A.A., Hermanson, L., Materia, S., Sanna, A., Borrelli, A., MacLachlan, C. and Gualdi, S. (2017) A multisystem view of wintertime NAO seasonal predictions. *Journal of Climate*, 30(4), pp.1461-1475.

Baker, L. H., L. C. Shaffrey, R. T. Sutton, A. Weisheimer and A. A. Scaife (2018) An intercomparison of skill and over/underconfidence of the wintertime North Atlantic Oscillation in multi-model seasonal forecasts, *Geophys. Res. Lett.*, doi:10.1029/2018GL078838

Bengtsson, L., U. Andrae, T. Aspelien, Y. Batrak, J. Calvo, W. de Rooy, E. Gleeson, B. Hansen-Sass, M. Homleid, M. Hortal, K. Ivarsson, G. Lenderink, S. Niemelä, K.P. Nielsen, J. Onvlee, L. Rontu, P. Samuelsson, D.S. Muñoz, A. Subias, S. Tijm, V. Toll, X. Yang, and M.Ø. Køltzow (2017) The HARMONIE–AROME Model Configuration in the ALADIN–HIRLAM NWP System. *Mon. Wea. Rev.*, 145, 1919–1935, https://doi.org/10.1175/MWR-D-16-0417.1

Bonavita, M., L. Isaksen and E. Hólm (2012) On the use of EDA background error variances in the ECMWF 4D-Var. *Q. J. R. Meteorol. Soc.*, 138, 1540–1559, doi:10.1002/qj.1899.

Bushuk, M. et al. (2017) Skillful regional prediction of Arctic sea ice on seasonal timescales. *Geophys. Res. Lett.* 44, 4953–4964, doi:10.1002/2017GL073155

Cavalieri, D.J., C.L. Parkinson, P. Gloersen and H.J. Zwally (1996, updated yearly). Sea ice concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS passive microwave data, version 1. Boulder, Colorado USA. *NASA National Snow and Ice Data Center Distributed Active Archive Center.* doi: 10.5067/8GQ8LZQVL0VL. (Accessed 20 February 2017)

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P. , Poli, P. , Kobayashi, S. , Andrae, U. , Balmaseda, M. A., Balsamo, G. , Bauer, P. , Bechtold, P. , Beljaars, A. C., van de Berg, L. , Bidlot, J. , Bormann, N. , Delsol, C. , Dragani, R. , Fuentes, M. , Geer, A. J., Haimberger, L. , Healy, S. B., Hersbach, H. , Hólm, E. V., Isaksen, L. , Kållberg, P. , Köhler, M. , Matricardi, M. , McNally, A. P., Monge-Sanz, B. M., Morcrette, J. , Park, B. , Peubey, C. , de Rosnay, P. , Tavolato, C. , Thépaut, J. and Vitart, F. (2011), The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.*, 137: 553-597. doi:10.1002/qj.828

Dirkson, A., W. J. Merryfield and A. H. Monahan (2018) Calibrated probabilistic forecasts of Arctic sea ice concentration. *In revision for J. Climate.*

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., & Rodrigues, L. R. L. (2013) Seasonal climate predictability and forecasting: status and prospects. *Wiley Interdisciplinary Reviews: Climate Change*, 4(4), 245–268. doi:10.1002/wcc.217

Docquier, D., J. P. Grist, M. J. Roberts, C. D. Roberts, T. Semmler, L. Ponsoni, F. Massonnet, D. Sidorenko, D. Sein, D. Iovino, T. Fichefet (2018) Impact of model resolution on Arctic sea ice and North Atlantic ocean heat transport. *In preparation for Climate Dynamics*

Drobot S. D., J. A. Maslanik, C. F. Fowler (2006) A long-range forecast of Arctic summer sea-ice minimum extent. *Geophys. Res. Lett.*, 33, L10501, doi:10.1029/2006GL026216

Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L., & Robinson, N. (2014). Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophys. Res. Lett.*, 41(15), 5620-5628.

Ferranti L, Magnusson L, Vitart F, Richardson DS. (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*., in press. doi: 10.1002/qj.3341

Goessling, H.F. et al. (2016) Predictability of the Arctic sea ice edge. *Geophys. Res. Lett.* 43, 1642-1650, doi:10.1002/2015GL067232

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.*, in press. doi:10.1002/qj.3242

Guemas, V., E. Blanchard-Wrigglesworth, M. Chevallier, J. J. Day, M. Déqué et al. (2016) A review on Arctic sea-ice predictability and prediction on seasonal to decadal time-scales. *Q. J. R. Meteorol. Soc.* 142: 546-561, doi: 10.1002/qj.2401

Hagedorn, R., F.J. Doblas-Reyes and T.N. Palmer (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting - I. Basic concept. *Tellus,* 57A, 219-233.

Hansen, B. B., Isaksen, K., Benestad, R. E., Kohler, J., Pedersen, Å. Ø., Loe, L. E., et al. (2014). Warmer and wetter winters: Characteristics and implications of an extreme weather event in the High Arctic. *Environ. Res. Lett.*, 9(11), 114021. https://doi.org/10.1088/1748-9326/9/11/114021

Køltzow, M., B. Casati, E. Bazile, T. Haiden and T. Valkonene (2018). A NWP model inter-comparison of surface weather parameters during the Year of Polar Prediction Special Observing Period 1, *in preparation.*

Lavers, D.A., E. Zsoter, D.S. Richardson, and F. Pappenberger (2017) An Assessment of the ECMWF Extreme Forecast Index for Water Vapor Transport during Boreal Winter. *Wea. Forecasting,* **32**, 1667–1674, doi:10.1175/WAF-D-17-0073.1

Leutbecher, M., and T. N. Palmer (2008) Ensemble forecasting. *J. Computational Physics*, 227, 3515–3539.

Leutbecher, M. et al. (2017) Stochastic representations of model uncertainties at ECMWF: state of the art and future vision. *Q. J. R. Meteorol. Soc.*, 143 (707), 2315–2339, doi:10.1002/qj.3094.

Lindsay R. W., J. Zhang, A. J. Schweiger, M. A. Steele (2008) Seasonal predictions of ice extent in the Arctic Ocean. *J. Geophys. Res.*, 113, C02023, doi:10.1029/2007JC004259

MacLachlan, C., A. Arribas, K. A. Peterson, A. Maidens, D. Fereday, A. A. Scaife, M. Gordon, M. Vellinga, A. Williams, R. E. Comer, J. Camp, P. Xavier and G. Madec (2015) Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system. *Q. J. R. Meteorol. Soc.*, 141, 1072-1084, doi:10.1002/qj.2396.

Magnusson, L. and E. Källén (2013) Factors Influencing Skill Improvements in the ECMWF Forecasting System. *Mon. Wea. Rev.,* **141**, 3142–3153, doi: 10.1175/MWR-D-12-00318.1

Melia N., K. Haines, E. Hawkins and J. J. Day (2017) Towards seasonal Arctic shipping route predictions. *Environ. Res. Lett.*, 12, 084005, doi: 10.1088/1748_9326/aa7a60

Mogensen, K. S., L. Magnusson, and J. Bidlot (2017) Tropical cyclone sensitivity to ocean coupling in the ECMWF coupled model. *J. Geophys. Res.: Oceans*, 122 (5), 4392–4412, doi:10.1002/2017JC012753.

Müller, M., Y. Batrak, J. Kristiansen, M.A. Køltzow, G. Noer, and A. Korosov (2017) Characteristics of a Convective-Scale Weather Forecasting System for the European Arctic. *Mon. Wea. Rev.*, 145, 4771–4787, doi: 10.1175/MWR-D-17-0194.1

Müller W. A., J. H. Jungclaus, T. Mauritsen, J. Baehr, M. Bittner, R. Budich, F. Bunzel, M. Esch, R. Ghosh, H. Haak, T. Ilyina, T. Kleine, L. Kornblueh, H. Li, K. Modali, D. Notz, H. Pohlmann, E. Roeckner,, I. Stemmler, F. Tian, J. Marotzke (2018) A higher-resolution version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR). *Journal of Advances in Modeling Earth Systems* 10:1383–1413, doi: 10.1029/2017MS001217

Parkinson, C. L., and Comiso J. C. (2013) On the 2012 record low Arctic sea ice cover: Combined impact of preconditioning and an August storm. *Geophys. Res. Lett.* 40, 1356–1361, doi: 10.1002/grl.50349.

Pauley, P.M., 1998: An Example of Uncertainty in Sea Level Pressure Reduction. *Wea. Forec.* **13**, 833–850, doi: 10.1175/1520-0434(1998)013<0833:AEOUIS>2.0.CO;2

Rabier, F., H. Jarvinen, E. Klinker, J. Mahfouf, and A. Simmons, 2000: The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Q. J. R. Meteorol. Soc.*, 126 (564), 1143–1170, doi:10.1002/qj.49712656415.

Roberts M. J. , et al. (2018a), Description of the resolution hierarchy of HadGEM3-GC3.1 model as used in HighResMIP coupled experiments. *Geosc. Mod. Dev.*

Roberts C. D., R. Senan, F. Molteni, S. Boussetta, M. Mayer, S. P. E. Keeley (2018b) Climate model configurations of the ECMWF Integrated Forecast System (ECMWF-IFS cycle 43r1) for HighResMIP. *Geosc. Mod. Dev.* 11:3681–3712, DOI 10.5194/gmd-11-3681-2018

Scaife, A. A., et al. (2014) Skillful long-range prediction of European and North American winters. *Geophys. Res. Lett.*, 41, 2514–2519, doi: 10.1002/2014GL059637.

Serreze, M. C., Crawford, A. D. and Barrett, A. P. (2015) Extreme daily precipitation events at Spitsbergen, an Arctic Island. *Int. J. Climatol.*, 35: 4574-4588. doi:10.1002/joc.4308

Sidorenko D., T. Rackow, T. Jung, T. Semmler, D. Barbi, S. Danilov, K. Dethloff, W. Dorn, K. Fieg, H. F. Goessling, D. Handorf, S. Harig, W. Hiller, S. Juricke, M. Losch, J. Schröter, D. V. Sein, Q. Wang

(2015) Towards multiresolution global climate modeling with ECHAM6FESOM. Part I: model formulation and mean climate. *Climate Dyn.* 44:757–780, DOI 10.1007/s00382-014-2290-6

Tibaldi, S. and Molteni, F. (1990) On the operational predictability of blocking, *Tellus A*, 42:3, 343-365, doi: 10.3402/tellusa.v42i3.11882

Vautard, R. (1990) Multiple Weather Regimes over the North Atlantic: Analysis of Precursors and Successors. *Mon. Wea. Rev.,* **118**, 2056–2081, doi: 10.1175/1520-0493(1990)118<2056:MWROTN>2.0.CO;2

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

Voldoire, A., D. Saint-Martin, S. Sénési, A. Alias et al. (2018) CNRM-CM6-0: description and validation. *In preparation for Clim. Dyn.*

Wang Q., S. Danilov, D. Sidorenko, R. Timmermann, C. Wekerle, X. Wang, T. Jung, J. Schröter (2014) The Finite Element Sea Ice-Ocean Model (FESOM) v.1.4: formulation of an ocean general circulation model. *Geosc. Mod. Dev.* 7:663–693, doi: 10.5194/gmd-7-663-2014

Williams K. D., D. Copsey, E. W. Blockley, A. Bodas-Salcedo, D. Calvert, R. Comer, T. Graham, H. T. Hewitt, R. Hill, P. Hyder, S. Ineson, T. C. Johns, A. B. Keen, R. W. Lee, A. Megann, S. F. Milton, J. G. L. Rae, M. J. Roberts, A. A. Scaife, R. Schiemann, D. Storkey, L. Thorpe, I. G. Watterson, D. N. Walters, A. West, R. A. Wood, T. Woollings, P. K. Xavier (2018) The Met Office Global Coupled Model 3.0 and 3.1 (GC3.0 and GC3.1) Configurations. *Journal of Advances in Modeling Earth Systems* 10:357–380, doi: 10.1002/2017MS001115

Woollings, T. , Hannachi, A. and Hoskins, B. (2010) Variability of the North Atlantic eddy‐driven jet stream. *Q.J.R. Meteorol. Soc.*, 136: 856-868. doi:10.1002/qj.625

Zuo, H., M. Alonso-Balmaseda, K. Mogensen, and S. Tietsche (2018) Ocean5: The ECMWF ocean reanalysis system and its real-time analysis component. *ECMWF Technical Memorandum* 823

Websites

https://titan.uio.no/node/2009 (in Norwegian, accessed August 2018).

https://www.dagbladet.no/nyheter/130-evakuert-pa-svalbard-deler-av-longyearbyen-er-sperret-av-og-det-er-innfort-ferdselsforbud/64354331 (in Norwegian, accessed August 2018).

https://norut.no/nb/news/kartlegging-med-radarsatellitt-gir-bedre-snoskredvarsling-og-beredskap (in Norwegian, accessed August 2018).

8. ACRONYMS

AEE: Absolute Extent Error

ACC: Anomaly Correlation Coefficient

CMIP6: Coupled Model Intercomparison Project phase 6

DJF: December-January-February

ECMWF: European Centre for Medium-range Weather Forecasts

EDA: Ensemble 4D-Var Assimilations

EFI: Extreme Forecast Index

ENS: ECMWF ensemble forecast

ENSO: El Niño Southern Oscillation

EOF: Empirical Orthogonal Function

ETS: Equitable Threat Score
FB: Frequency Bias
FCRPSS: Fair Continuous Ranked Probability Skill Score
GCM: Global Coupled Model
GPCP: Global Precipitation Climatology Project
HARMONIE: HIRLAM-ALADIN Research on Mesoscale Operational NWP in Europe
HIRLAM: High Resolution Limited Area Model
IFS: Integrated Forecasting System
IIEE: Integrated Ice Edge Error
IVT: Integrated water Vapour Transport
JJA: June-July-August
MAE: Mean Absolute Error
MC: Monte Carlo
ME: Mean Error (sections 3 and 4) or Misplacement Error (section 5.2)
MME: Multi-model ensemble
MSLP: Mean Sea-Level Pressure
NAO: North Atlantic Oscillation
NEMO: Nucleus for European Modelling of the Ocean
NSIDC: NASA National Snow and Ice Data Center
NWP: Numerical Weather Prediction
OHT: Ocean Heat Transport
PRIMAVERA: H2020 project Process-based climate sIMulation: AdVances in high-resolution modelling and European climate Risk Assessment
RMSE: Root Mean Square Error
RPC: Ratio of Predictable Components
SEAS5: ECMWF Seasonal forecasting system 5
SIA: Sea Ice Area
SIC: Sea Ice Concentration
SIE: Sea Ice Extent
SIT: Sea Ice Thickness
SIV: Sea Ice Volume
SOP1: First Special Observing Period (of YOPP)
SPS: Spatial Probability Score
SST: Sea Surface Temperature
TCC: Total Cloud Cover
UM: UK Met Office Unified Model
YOPP: Year Of Polar Prediction

9. ANNEXES

9.1. Winter sea ice extent seasonal re-forecasts

This annex presents and discusses results for the seasonal re-forecasts evaluated in section 5.2 for the November initialization, focusing on winter and the March maximum of the annual cycle of sea ice extent.

Figure 9.1.1 presents the evolution of RMSE and correlation according to the forecast month for re-forecasts initialized in November. Unlike results for the summer season (May starts), very minor differences between the different systems are found, and levels of skill are clearly better than that of persistence. These results are consistent with previous findings which established a lower predictability in forecasts initialized in late spring (e.g. Guemas et al. 2016).

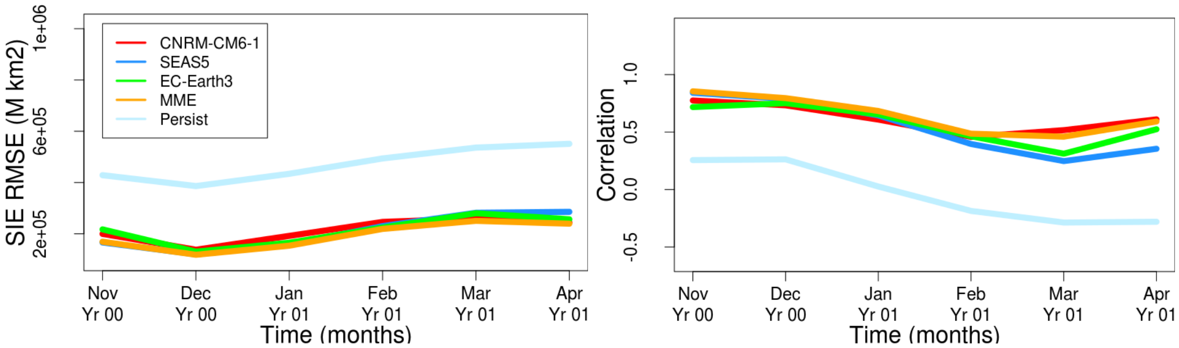


Fig. 9.1.1 Evolution according to forecast month of pan-Arctic SIE RMSE (left) and correlation (right) with NSIDC reference data in re-forecasts initialized in November 1993-2014. The multi-model ensemble (MME) is shown in orange, and persistence of April anomalies in light blue.

Figure 9.1.2 shows the IIEE and decomposition for each system as well as for the multi-model ensemble. As for the May starts, no significant improvement with a simple multi-model approach is found, possibly due to the strong similarities between the different forecasting systems. The decomposition in ME and AEE shows a much higher inter-annual variability for the winter season than for the summer season. This is possibly mainly due to the fact that ice-free areas are more restricted during winter. As for the probabilistic score (SPS, fig. 9.1.3), very minor differences between the systems are found. The highest differences appear in the very first month of the re-forecast, again suggesting the importance of initialization in the forecast quality.

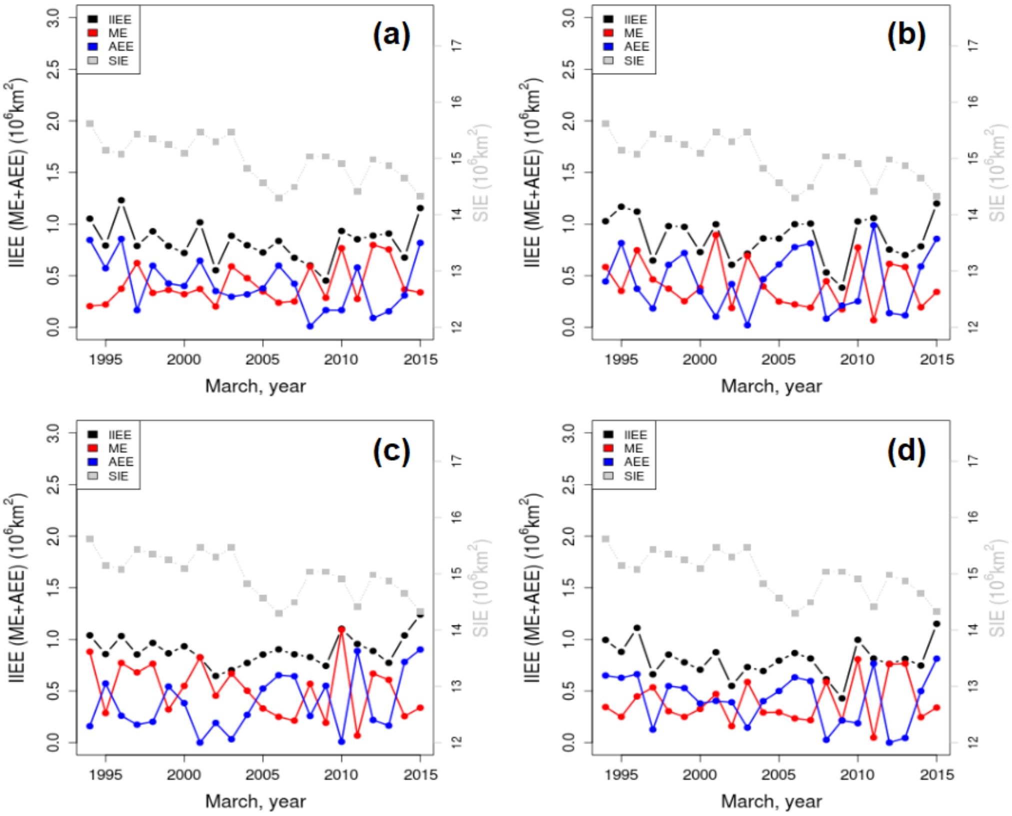


Fig 9.1.2: IIEE (black, in millions of km2) and decomposition in ME (red) and AEE (blue) with respect to NSIDC data for March 1993 to 2014 in re-forecasts initialized in November with (a) CNRM-CM6-1, (b) SEAS5 and (c) EC-Earth3. (d) Same as (a-c) but for a multi-model ensemble grouping all ensemble members of each individual system (after individual bias correction of SIC). The grey line shows the reference SIE (y-axis on the right hand side).

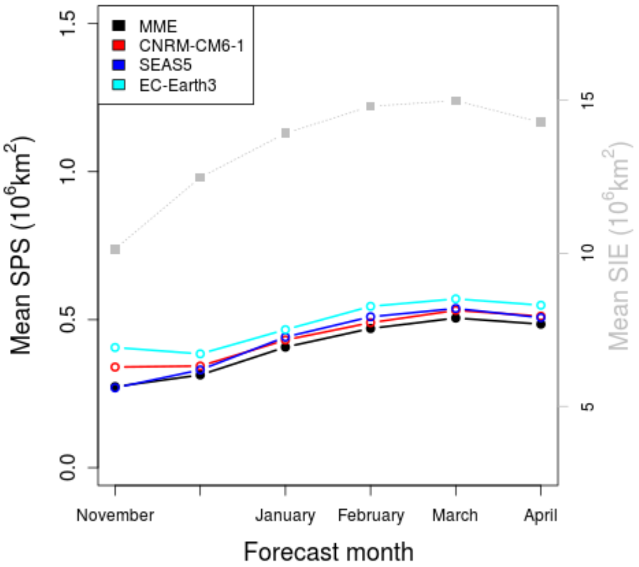


Fig. 9.1.3 Mean SPS over 1993-2014 according to forecast month for each system and the MME (in black) for re-forecasts initialized in November.