THEME ENV.2012.6.1-1

G.A. no 308378

WP 3.1 Impact of improved initialisation and sample size
Deliverable D31.1 Forecast quality assessment of the backward-extended decadal hindcast experiments

Date: 30/04/2015

Deliverable Title	Forecast quality assessment of the backward-extended decadal hindcast experiments			
WP number and title	WP 3.1 Impact of improved initialisation and sample size			
Lead Beneficiary	*MPG*			
Contributors	*Wolfgang Müller, Holger Pohlmann, Frank Sienz*	*MPG*		
	Wilco Hazeleger, Camiel Severijns	*KNMI*		
	Scientists involved	*Organisation name*		
	Scientists involved	*Organisation name*		
Creation Date	12/03/2015	**Version number**	4	
Deliverable Due Date	30/04/2015	**Actual Delivery Date**	04/05/20145	
Nature of deliverable	x	R - Report		
		P - Prototype		
		D - Demonstrator		
		O – Other		
Dissemination Level/ Audience		PP - Public		
	x	PU - Restricted to other programme participants, including the Commission services		
		RE - Restricted to a group specified by the consortium, including the Commission services		
		CO - Confidential, only for members of the consortium, including the Commission services		

Version	Date	Modified by	Comments
1	12/03/2015	Wolfgang Müller	
2	15/04/2015	Francisco J. Doblas-Reyes	
3	21/04/2015	Camiel Severijns	
4	22/04/2015	Wolfgang Müller	

INDEX

1. Executive summary

Initialization of coupled climate models from near observational states has been shown to improve decadal climate predictions. In the fifth coupled model intercomparison project (CMIP5) improved skill is found particularly in the North Atlantic and Pacific (Doblas-Reyes et al. 2013). However, these previous assessments are restricted by observations and reanalyses, which cover the period from 1960 to the present day. Given the short period, the skill assessment considers only a small number of cycles of the observed decadal to multidecadal variability and is furthermore strongly affected by the trend. However, variability around the trend becomes more important over longer periods. This can have an effect on the forecast skill estimation. Here we initialize a conceptual model to systematically assess the effect of sample size (ensemble size and hindcast length) on to the North Atlantic SST forecast skill. The coupled Max Planck Institute Earth System Model (MPI-ESM) is applied to assess the global forecast skill in a set of retrospective predictions covering an extended period from 1901 to 2010. Further, a set of decadal predictions with the EC-Earth model version 2.3 have been completed for yearly start dates from 1920 to 1950.

The main results are that for small ensemble sizes (i) the root-mean squared error (RMSE) confidence intervals of North Atlantic SST are biased to too high values and (ii) the power of the performed statistical tests is low, while it depends also on the noise ratio. In addition, (iii) small ensembles can be compensated by large hindcast samples to yield the same test power, and vice versa. For the MPI-ESM, a backward extended hindcast period leads to an enlargement of regions with significant anomaly correlation coefficients (ACC) for predicted surface temperatures. This arises from a more realistic contribution of the trend, which is also found in the uninitialized runs. Additionally, in the North Atlantic decadal variability plays a larger role over the extended period, with detrended time series showing higher ACC for the extended compared to the short period.

2. Project objectives

With this deliverable, the project has contributed to the achievement of the following objectives (see DOW Section B.1.1.2):

No.	Objective	Yes	No
1.	To achieve an objective exhaustive *evaluation* of current forecast quality from dynamical, statistical, and consolidated systems to identify the factors limiting s2d predictive capability	X	
2.	To test specific hypotheses for the improvement of s2d predictions, including novel mechanisms responsible for high-impact events using a *process-based verification* approach		X
3.	To develop innovative methods for a comprehensive *forecast quality assessment*, including the maximum skill currently attainable	X	
4.	To facilitate the *integration of multidimensional observational data* of the atmosphere-ocean-cryosphere-land system as sources of initial conditions, and to validate and calibrate climate predictions		X
5.	To achieve an *improved forecast quality at regional scales* by better initialising the different components, an increase in the		X

No.	Objective	Yes	No
	spatial resolution of the global forecast systems and the introduction of important new process descriptions		
6.	To assess the best alternatives to characterise and deal with the *uncertainties in climate prediction* from both dynamical and statistical perspectives for the increase of forecast reliability	X	
7.	*To achieve reliable and accurate local-to-regional predictions* via the combination and calibration of the information from different sources and a range of state-of-the-art regionalisation tools		X
8.	*To illustrate the usefulness* of the improvements for specific applications and develop methodologies to better communicate actionable climate information to policy-makers, stakeholders and the public through peer-reviewed publications, e-based dissemination tools, multi-media, examples for specific stakeholders (energy and agriculture), stakeholder surveys, conferences and targeted workshops		X
9.	*To support* the European contributions to *WMO research initiatives* on s2d prediction such as the GFCS and enhance the European role on the *provision of climate services* according to WMO protocols by creating examples of improved tailored forecast-based products for the GPCs and participating in their transfer to worldwide RCCs and NHMSs.		X

3. Detailed report on the deliverable

Hindcast retrospective prediction experiments have to be performed to validate decadal prediction systems. These are necessarily restricted in their number due to both observational and computational constrains. From weather and seasonal prediction it is known that the ensemble size is crucial (e.g. Scaife et al. 2014). The hindcast length, however, also affects the forecast skill assessment as shown for seasonal timescales (Müller et al. 2005, Shi et al. 2015). A similar dependency is likely to be found for decadal predictions although differences are expected due to the differing time scales of the processes involved and the longer prediction horizon.

It is shown here, that the ensemble and hindcast sample sizes have a large impact on the uncertainty assessment of the ensemble mean, as well as for the detection of prediction skill. For this purpose a conceptual model that enables the systematic analysis of statistical properties and its dependencies in a framework close to that of real decadal predictions is developed (Sienz et al. 2015). In addition, a set of extended range hindcast experiments have been undertaken, covering the entire 20th century (Müller et al. 2014).

The conceptual model

An analysis of confidence interval (CI) and statistical test properties requires the knowledge of the true values. For that reason, an idealised conceptual model that includes the basic components associated with decadal predictability has been created (Sienz et al. 2015). The components used for the

simulation study are a nonlinear trend (a regression on global CO2 concentration), a multi-decadal oscillation with a period of 70 years and a Gaussian white noise. These components together are compared with the observed North Atlantic sea surface temperature (NA-SST). Simulations of initialised decadal predictions (Isim) and ”observations” (details in the figure captions) include all components and differ only in the noise term. The phase of the multi-decadal oscillation is randomly perturbed to generate simulated uninitialised predictions (Usim). The analysis below requires a large number of repetitions, therefore 10,000 simulations are performed.

Properties of confidence intervals and tests

The quality of CIs is determined by their coverage and imbalance. The number of times the CI includes the true value is the coverage, whereas the imbalance measures the percentage of CI above or below the true value. These properties are calculated for the CI of the root-mean-squared error (RMSE) of the ensemble mean. The RMSE is chosen because it is a stronger (more general) criterion than for example the correlation. Uncertainty is assessed with the recommended resampling procedure (Goddard et al. 2013). For small ensemble sizes the coverage is low (Fig. 1a) and the RMSE CI are strongly biased to too high RMSE (Fig. 1b). This imbalance is also present for larger ensemble sizes (Fig. 1b).

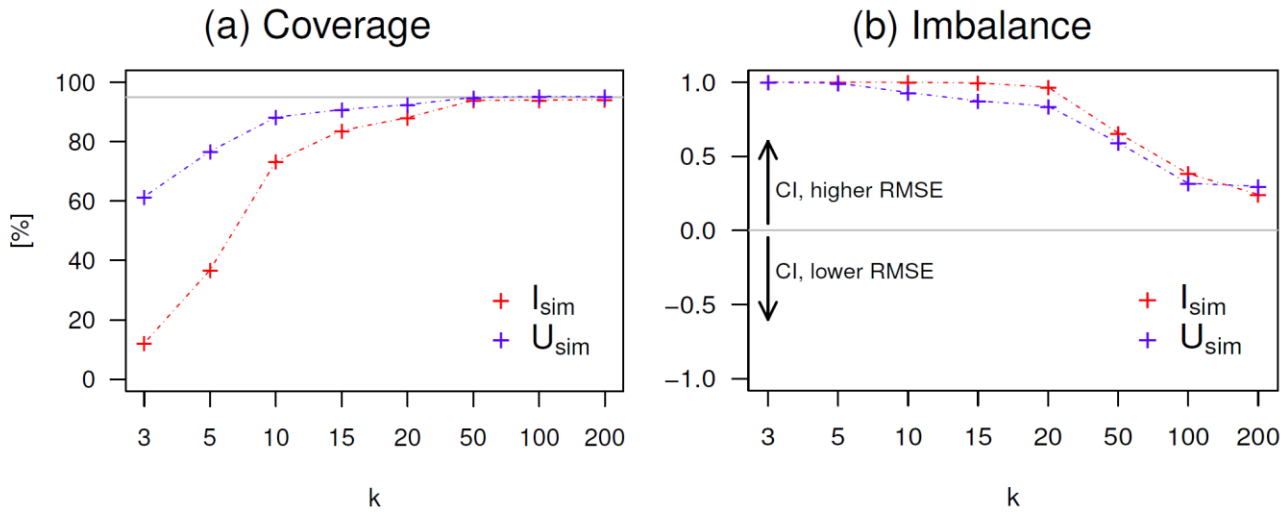

Fig. 1. (a) Coverage and (b) imbalance of the 95% confidence intervals (CI) for the RMSE of the ensemble mean simulated initialised (Isim) and uninitialised (Usim) predictions. K denotes the ensemble size.

The performance of significance tests is investigated in terms of power. The power depends on the size of the effect, the noise ratio (the variance of the unpredictable noise component), as well as on the sample size. Here, the detectability of a positive RMSE skill score of Isim is investigated, taking Usim as the reference prediction. This is an analogue situation to the detection of an improvement by providing starting conditions for decadal prediction systems. Only low power can be achieved with small ensemble sizes (Fig. 2a).

The power is even lower for short hindcast sample sizes or with an increased noise ratio (Fig. 2b). In addition, the required ensemble size to achieve a power of 80% for a given hindcast sample size is calculated. This is expressed in computation time needed to perform the hindcast experiment and is represented by model (simulation) years. The model years are given by the product of the number of required ensemble members, initialisations and years predicted, assuming yearly initialisation and 10 year predictions (Fig. 2c). For the presented example hindcasts started in 1940 require only 15

ensemble members compared to 52 members when started in 1960 to achieve a power of 80%. This corresponds to a reduction of model years by more than half.

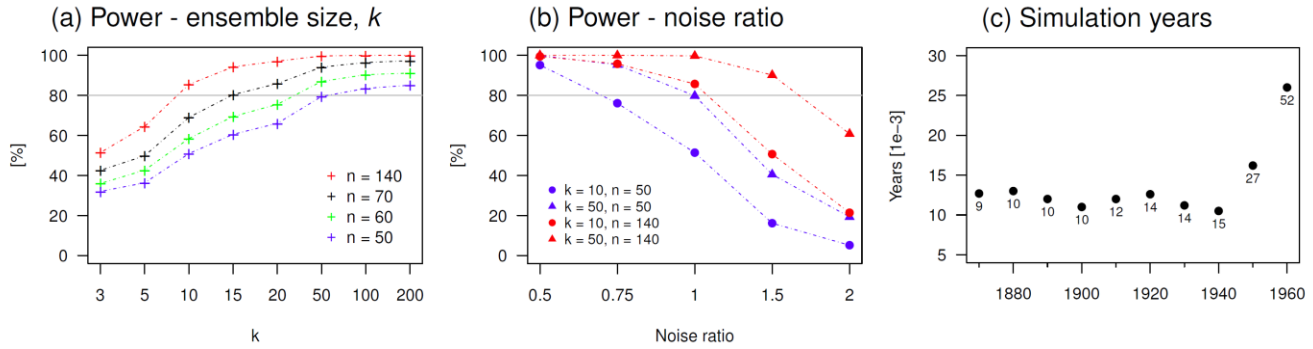

Fig. 2. Statistical power to detect an RMSE skill score of 0.3 as significantly greater than zero for (a) increasing ensemble size, k, and (b) different random noise ratios. n is the hindcast sample size. (c) Model years needed to achieve a power requirement of at least 80% in dependence on the hindcast sample size. The years given at the horizontal axis indicates the beginning of the hindcast period. The numbers are the corresponding ensemble sizes.

Decadal predictions and ensemble size

The resampling scheme for uncertainty assessment can be extended to account for different ensemble sizes in climate prediction experiments (Sienz et al. 2015). This is exemplified here for the MiKlip Baseline1 predictions (see Pohlmann et al. 2013) of the NA-SST and the central Europe summer temperatures (CEU-JJA) with 10 ensemble members and yearly initialisation from 1960 to 2009. Correlation increases and its uncertainty range decreases with larger ensemble sizes, consequently p-values for the correlation skill scores (css; given by (C(init) - C(uninit))/(1 - C(uninit))) yield smaller values (Fig. 3a). The NA-SST correlations stabilize quickly (Fig. 3a) in contrast to the ones of CEU-JJA (Fig.3b). According to the theoretical approximation by Murphy (1990), the CEU-JJA css saturates with more than 50 members and reaches a value 0.6.

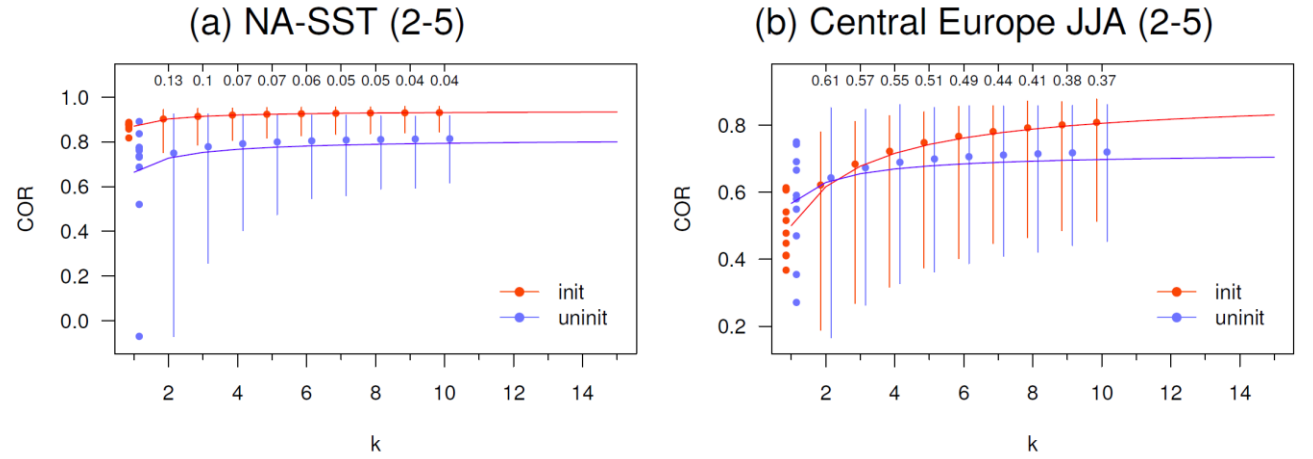

Fig. 3. Correlation and 95% CI for (a) NA-SST and (b) CEU-JJA as a function of the ensemble size for initialized predictions (init) and uninitialised historical runs (uninit) and prediction years 2 to 5. Numbers at the top are p-values for the correlation skill scores of init, with the reference prediction uninit. The null hypothesis is, that the css is equal to zero. The lines are theoretical approximations (Murphy 1990).

Extended hindcast experiments for the 20th century

Retrospective predictions back to 1901 have been performed with the MPI-ESM coupled climate model. For this purpose, three assimilation runs were performed based on a reconstructed ocean state with MPIOM (Müller et al. 2014a). Based on this, three hindcast experiments are completed with a yearly initialization from 1900 to 2009. The initialised predictions of the NA-SST outperform the uninitialised runs and reproduce the higher temperatures observed in 1930s and 1940s, as well as the cold phase during the 1970s (Fig. 4a). Consequently, the detrended series of the initialised predictions result in higher correlations, not only for years 2 to 5, but also for higher lead times (Fig. 4b). The extended hindcast further reveals an enlargement of significant surface temperature correlations at a global scale in comparison to the reduced time period (1960 - 2009; Fig. 5), which is the standard in real-time decadal prediction. The output data of the backward extended decadal predictions has been CMOR-ized and uploaded to the SPECS data archive.

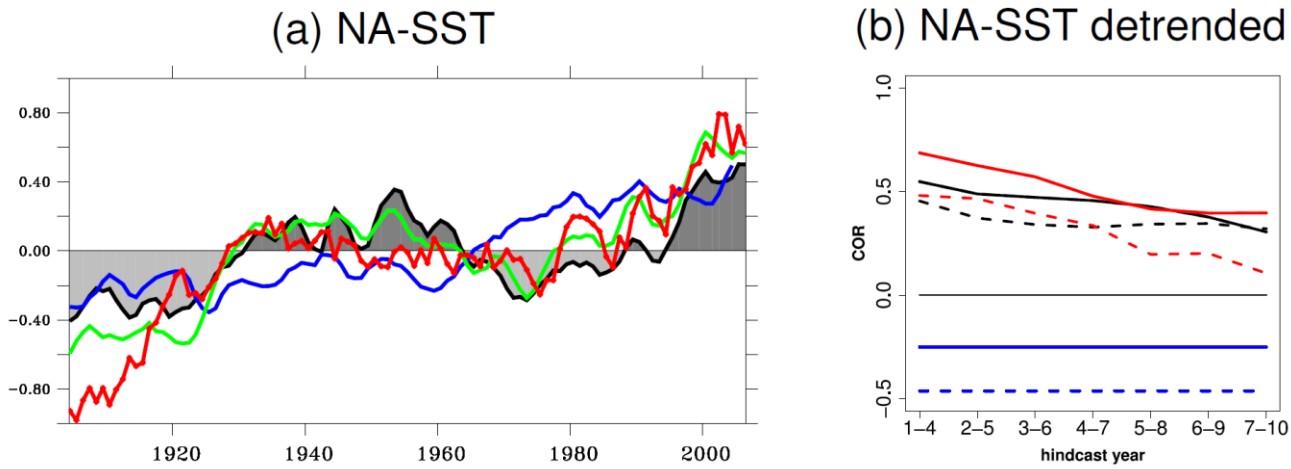

Fig. 4. (a) Four-year means of NA-SST; ensemble means of the 20th century reanalysis (20CR; black), assimilation run (green), the uninitialised runs (blue) and retrospective predictions (red) are shown. (b) Correlations with 20CR for the detrended four-year mean time series, persistence predictions (black), solid and dashed lines are the correlations for the backward extended time period (1900-2009) and the reduced time period (1960-2009), respectively.

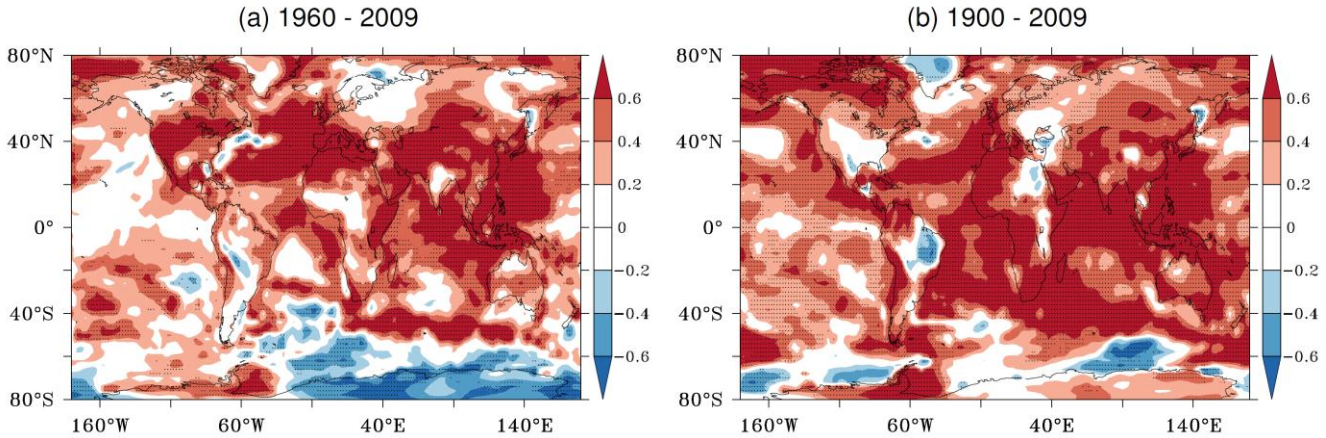

Fig. 5. Correlation between the ensemble mean surface air temperature predictions (years 2-5) and 20CR for the (a) reduced and (b) complete time period. Stippled areas indicate significant values at a 95% confidence level.

Further, a set of decadal predictions with the EC-Earth model version 2.3 have been completed for yearly start dates from 1920 to 1950. For each start date an ensemble of 12 members was started on November 1 of the year before and run for 10 years. The initial conditions for the ocean were created by forcing the ocean model from EC-Earth with the atmospheric boundary layer state of 12 members of the 20th Century Reanalysis data set (20th Century Reanalysis V2 data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at http://www.esrl.noaa.gov/psd/). This forced ocean model ensemble was run from 1870 to 1960.

By starting the forced ensemble 50 years earlier than the time period of the decadal predictions ensures that there is sufficient spread in the ocean initial conditions. The atmospheric initial conditions were taken from a 16 member ensemble of EC-Earth runs. This ensemble of coupled atmosphere-ocean runs was forced with CMIP5 historical forcing data from 1850 to 2005. The output data of the decadal predictions has been CMORized and uploaded to the CEDA data archive at BADC. At this moment most of this data has been published as well on the SPECS data portal at CEDA.

In Fig. 6 the anomaly of the ensemble and global mean of the sea surface temperature is shown for all decadal predictions performed with the EC-Earth model. In Fig. 6 the anomaly of the ensemble mean of the NA-SST is shown for all decadal predictions performed with the EC-Earth model. The decadal predictions were bias corrected using the mean drift of all decadal predictions with respect to the mean of the CMIP5 historical ensemble.

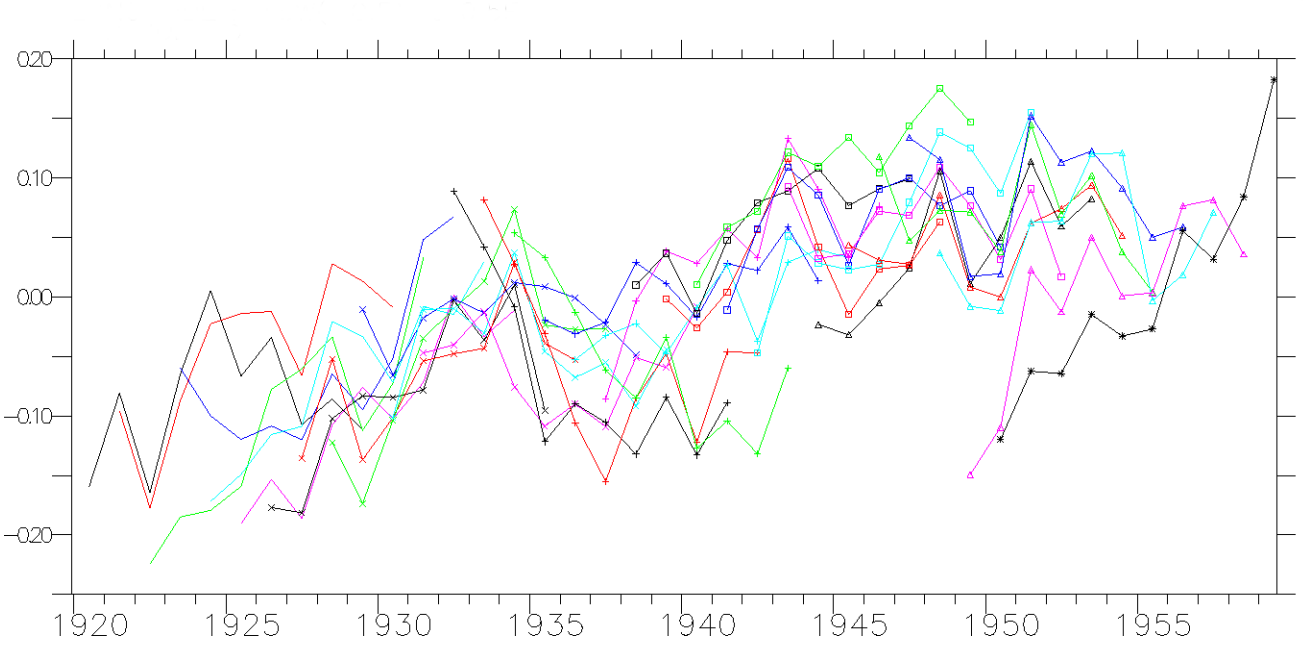

Fig. 6. The ensemble mean sea surface temperature anomaly in the North Atlantic for all decadal predictions with EC-Earth. The predicted anomalies are plotted in a different color and line style for each start data. The mean drift of all decadal predictions with respect to a sixteen member EC-Earth ensemble with CMIP5 historical forcing was used as a bias correction.

4. Conclusions and discussion

Small ensemble and hindcast sample sizes lead to biased inference test performances in a way that the detection of a present prediction skill is hampered. This is not only relevant for the validation of decadal prediction systems, but also for their development. For example, the comparison of different starting conditions, initialisation strategies or methods for the ensemble generation, to name a few, is

made difficult and misleading conclusions can result in the presence of a low statistical power. From this point of view, the computational resources are a limiting factor for the validation and improvement of decadal prediction systems. On the other hand, it is worth to emphasize that the presented analysis suggests that skill beyond the one presented is already present in the actual prediction systems.

Moreover, for the backward extended hindcast experiments based on the coupled model, a careful examination of the 20th century data (20Cr or ERA-CLIM) assimilation experiments are required. For example in the MPI-ESM experiments, the assimilated and initialized SST anomalies have particular large negative temperature anomalies during the first two decades of the experiments (Fig. 4a), which are due to a strong cooling of the North Atlantic of the forced ocean model in the first decades (Müller et al. 2014a). Here the 20CR forced ocean model and subsequently the assimilation runs exhibit a strong intrusion of Arctic freshwater into the North Atlantic, leading to cooler and fresher water masses particularly in the SPG region until the 1910s. Hence to use this assimilation to initialize decadal prediction up to the beginning of the 20th century this problem needs to be fixed.

5. References

- Goddard, L. et al. (2013), A verification framework for interannual-to-decadal predictions experiments, Clim. Dyn.,40, 245-272.
- Doblas-Reyes, F. J., I. Andreu-Burillo, Y. Chikamoto, J. García-Serrano, V. Guemas, M. Kimoto, T. Mochizuki, L. R. L. Rodrigues, and G. J. van Oldenborgh (2013), Initialized near-term regional climate change prediction, Nat. Commun., 4, 1715, doi:10.1038/ncomms2704.
- Scaife, A. A., A. Arribas, E. B Lockley, A. B Rookshaw, R. T. Clark, N. Dunestone, R. Eade, D. Fereday, C. K. Folland, M. Gordon, L. Hermanson, J. R. Knight, D. J. Lea, C. MacLachlan, A. Maidens, M. Martin, A. K. Peterson, D. Smith, M. Vellinga, E. Wallace, J. Waters, A. Williams (2014), Skillful long-range predictions of European and North American winters. Geophys. Res. Lett. 41, 2514–2519.
- Shi, W., N. Schaller, D. MacLeod, T.N. Palmer and A. Weisheimer (2015), Impact of hindcast length on estimates of seasonal climate predictability http://onlinelibrary.wiley.com/doi/10.1002/2014GL062829/references
- Murphy, J. M. (1990), Assessment of the practical utility of extended range ensemble forecasts. Q. J. R. Meteorol. Soc.,116, 89-125.
- Müller, W. A., C. Appenzeller and C. Schär (2005), Probabilistic Seasonal Prediction of the Winter North Atlantic Oscillation and its Impact on Near Surface Temperature, Clim. Dyn., 24, 213-226.
- Müller W. A., D. Matei, M. Bersch, J. H. Jungclaus, H. Haak, K. Lohmann,G. P. Compo, and J. Marotzke (2014a), A 20th-century reanalysis forced ocean model to reconstruct North Atlantic climate variation during the 1920s, Clim. Dyn.. doi:10.1007/s00382-014-2267-5.
- Pohlmann, H., W. A. Müller, K. Kulkarni, M. Kameswarrao, D. Matei, F. S. E. Vamborg, C. Kadow, S. Illing, and J. Marotzke (2013), Improved forecast skill in the tropics in the new MiKlip decadal climate predictions. Geophys. Res. Lett., 40, 5798-5802.

6. List of publications

Peer reviewed articles:

- Müller, W. A., H. Pohlmann, F. Sienz and D. Smith (2014b), Decadal climate prediction for the period 1901-2010 with a coupled climate model. Geophys. Res. Lett., 41, 2100-2107.

- Sienz, F., H. Pohlmann and W. A. Müller (2015), Ensemble size impact on the decadal predictive skill assessment. Met. Zeitschriften (submitted MetZ)

Plan for future publication:

None.

7. Efforts for this deliverable

Partner	Person-months (actual)	Person-months (in-kind)	Period covered
MPG	2		M1 -M36
KNMI	6		M1 -M36
IC3	3		M1 -M36
Total	**11**		