Seasonal Forecasts of Winter Temperature Improved by Higher-Order Modes of Mean Sea Level Pressure Variability in the North Atlantic Sector

Mikhail Dobrynin[1,1], Clementine Dalelane[1,1], and Kristina Fröhlich[1,1]

[1]Deutscher Wetterdienst

November 30, 2022

Abstract

The variability of the sea level pressure in the North Atlantic sector is the most important driver of weather and climate in Europe. The main mode of this variability, the North Atlantic Oscillation (NAO), explains up to 50% of the total variance. Other modes, known as the Scandinavian index, East Atlantic and East Atlantic/West Russian pattern, complement the variability of the sea level pressure, thereby influencing the European climate. It has been shown previously that a seasonal prediction system with enhanced winter NAO skill due to ensemble subsampling entails an improved prediction of the surface climate variables as well. Here, we show that a refined subselection procedure that accounts both for the NAO index and for the three additional modes of sea level pressure variability, is able to further increase the prediction skill of wintertime mean sea level pressure, near-surface temperature and precipitation across Europe.

Seasonal Forecasts of Winter Temperature Improved by Higher-Order Modes of Mean Sea Level Pressure Variability in the North Atlantic Sector

Clementine Dalelane, Mikhail Dobrynin, Kristina Fröhlich

Deutscher Wetterdienst (DWD)

Key Points:

- The hybrid seasonal forecast combines a dynamical forecast ensemble and a statistical prediction of general atmospheric circulation indices
- Dynamical seasonal forecasts are subsampled with respect to four statistically predicted circulation indices in the North Atlantic sector
- Forecast skill of sea level pressure, surface temperature and precipitation is improved across Europe compared to ensemble mean forecasts

Corresponding author: Clementine Dalelane, `clementine.dalelane@dwd.de`

Abstract

The variability of the sea level pressure in the North Atlantic sector is the most important driver of weather and climate in Europe. The main mode of this variability, the North Atlantic Oscillation (NAO), explains up to 50% of the total variance. Other modes, known as the Scandinavian index, East Atlantic and East Atlantic/West Russian pattern, complement the variability of the sea level pressure, thereby influencing the European climate. It has been shown previously that a seasonal prediction system with enhanced winter NAO skill due to ensemble subsampling entails an improved prediction of the surface climate variables as well. Here, we show that a refined subselection procedure that accounts both for the NAO index and for the three additional modes of sea level pressure variability, is able to further increase the prediction skill of wintertime mean sea level pressure, near-surface temperature and precipitation across Europe.

Plain Language Summary

Atmospheric winter conditions in Europe are primarily controlled by the varying pressure field over the North Atlantic, inducing generally cold/mild and dry/wet weather in Europe. Current seasonal forecasts of European winter climate, though highly desirable for society and economy, are as yet not fully reliable. There exist a number of autumn predictors, such as sea surface and stratospheric temperature, Eurasian snow depth, and Arctic sea ice, that impact on the upcoming pressure regimes in a predictable way. The present dynamical seasonal forecast systems respond still too weakly to these known seasonal predictors. But the relationship is reproduced quite well by means of statistics. In combination, statistical and dynamical forecasts have the potential to improve forecasts of the North Atlantic pressure conditions and thereby affected variables like temperature and precipitation in Europe considerably. We extend an existing hybrid seasonal forecast procedure by considering more modes of variability of the Atlantic pressure regimes than just the North Atlantic Oscillation. In this way, we are able to improve the forecasts for temperature and precipitation over wider regions in Europe.

1 Introduction

Seasonal prediction is a field of active research with several meteorological institutions worldwide issuing such seasonal forecasts to support environmental and economic decisions of a wide range of user groups. To date, the greatest success of such dynamical ensemble forecast systems is the prediction of ENSO (El Niño Southern Oscillation) several months ahead, which is the most important mode of interannual variability of the global climate influencing atmospheric phenomena around the world. In general, the skill of seasonal forecasts is satisfactory in the tropics, whereas prediction of northern mid-latitude seasonal climate remains challenging, as recently evaluated by Baker, Shaffrey, Sutton, et al. (2018). They show that the anomaly correlation coefficient (ACC) used to measure the prediction skill of mean sea level pressure (SLP) in a multi-model ensemble is low and not significant over most of the North Atlantic-European sector in most of the analyzed models.

Cohen et al. (2019) argue that new statistical techniques can increase the accuracy of seasonal forecasts and advocate the development of hybrid dynamical-statistical forecasts to produce more robust seasonal predictions. Hybrid forecasts based on circulation specification were presented for example by Baker, Shaffrey, and Scaife (2018) and Dobrynin et al. (2018).

In boreal winter, European weather and climate is dominated by the zonal propagation of planetary and synoptic-scale waves. This large scale circulation is an extremely high-dimensional phenomenon in real space. The technique of Principal Component Analysis (PCA), applied to the evolving sea level pressure (SLP) field, is one way to describe the states of this phenomenon in a sparse manner. The first principal component (PC) of SLP corresponds closely to the North Atlantic Oscillation (NAO) index, the importance of which for wintertime temperature, wind and precipitation anomalies in the North Atlantic-European sector has been known for long time (J. W. Hurrell, 1995; J. Hurrell et al., 2003; Thompson et al., 2003). However, despite its importance, it would be misleading to consider the NAO in isolation. Although PCs are orthogonal by construction, the components are interwoven nonlinearly, and every PC represents just one aspect of the whole circulation.

We therefore extend our notion of SLP variability considering three further modes of variability (2nd, 3rd and 4th PC) in addition to the NAO index. These modes, hence called circulation indices, correspond to the Scandinavian Index (SCAN), the East Atlantic/West Russian (EA/WR) and the East Atlantic (EA) pattern (although the denomination differs between authors, (Barnston & Livezey, 1987)). Together these indices explain about 80% of SLP variability. We set aside the inclusion of even more circulation indices, as their identification in short time series is complicated by stochastic noise.

Comas-Bru and McDermott (2014) show that higher-order circulation indices modulate the relation between NAO and European climate by shifting the NAO dipole in the South-West/North-East direction or rotating it in a clockwise/anticlockwise movement. Moreover, Vihma et al. (2018) explore the effects of large scale atmospheric patterns besides NAO on European winter temperatures.

Dobrynin et al. (2018) reported significant improvements in the seasonal prediction of surface temperature (TAS) and precipitation (PR) over a large area mostly in northern Eurasia: on the basis of an accurate prediction of the NAO index, "good" dynamical forecast members are selected from the forecast ensemble. But as the NAO index explains no more than 50% of the SLP variance, even a perfect prediction of the winter NAO will not improve the seasonal prediction of temperature and precipitation beyond certain limits (Dobrynin et al., 2018). The objective of the present paper is to explore possible improvements facilitated by the specification of all four leading circulation indices in the Euro-Atlantic sector (NAO, SCAN, EA/WR, EA).

To produce the mentioned accurate prediction of the NAO index, Dobrynin et al. (2018) developed a statistical estimator of the mean winter NAO index with a correlation of around 0.8 by taking into account autumn states of slowly varying boundary conditions of the ocean and atmosphere: arctic sea ice thickness, sea surface temperature, snow depth in Eurasia and stratospheric temperature in 100 hPa, see also Hall et al. (2017) and L. Wang et al. (2017). Similarly, Iglesias et al. (2014) and Ossó et al. (2018) predict the seasonal evolution of the East Atlantic pattern based on sea surface temperature. Rust et al. (2015) identify a linear relationship between temperature in Europe and several circulation indices, which allows the isochronic prediction of temperature anomalies given those indices.

We are going to broaden the approach of Dobrynin et al. (2018) by including the above mentioned predictor fields in four multiple linear regressions to predict each of the four considered circulation indices. These fields have been corroborated as physically meaningful drivers of the Euro-Atlantic SLP variability independently using causal network methods by Kretschmer et al. (2016). We show that an ensemble selection technique similar to Dobrynin et al. (2018), applied to the hindcasts of the operational seasonal forecast model of the German Meteorological Service GCFS2.0, accounting for four circulation indices, leads to substantial improvement in the forecasts of SLP, TAS and PR in the North Atlantic-European sector.

2 Data

We use data from the operational German Climate Forecast System, version 2 (GCFS2.0). GCFS2.0 is based on the MPI-ESM-HR (Müller et al., 2018; Mauritsen et al., 2018) with a horizontal resolution corresponding to 0.9° in the atmosphere and an ocean resolution of nominally 0.4°. In cooperation, Universität Hamburg (UHH), Max Planck Institute for Meteorology (MPI) and Deutscher Wetterdienst (DWD) have developed the seasonal prediction system GCFS, issuing operational seasonal forecasts once a month since 2016, starting on the first day of each month covering the upcoming 6 months. The first month is discarded as spin up.

The forecasts (both restrospective and real-time) are initialized with the state of the climate system inferred from the assimilation run using a continuous full-field nudging for ocean, sea-ice and atmosphere (Baehr et al., 2015). ERA-Interim vorticity, divergence, temperature and sea level pressure are used for the atmosphere, ORAS5 sea-ice, temperature and salinity are used for the ocean and sea-ice model. In order to account for uncertainties in initial conditions, an ensemble is established consisting of 50 members.

For each of the twelve forecasts per year, a hindcast data set (retrospective forecasts) consisting of 30 members per start date is provided to derive the model climate, error metrics and skill scores. In GCFS2.0, hindcast data cover the monthly starting dates from 1990 through 2017. The present study concentrates on hindcasts starting in November, which is when the upcoming boreal winter (December, January, February; DJF) is routinely forecasted.

As a complement to the assimilation run of the GCFS2.0 seasonal forecast system, we will also need the assimilation of the decadal prediction system developed in the MiKlip project (Pohlmann et al., 2019) because it extends 20 years farther into the past (1958-present). This system facilitates a slightly different initialization method compared to the seasonal prediction system. The atmosphere is nudged with ERA40 reanalysis full-field data until 1979 and ERA-Interim reanalysis data from 1980 onwards. The ocean is nudged with ORAS4 reanalysis anomalies during the whole duration (1960-present) of the simulation. The sea-ice is nudged with NSIDC sea-ice concentration anomalies from 1980 till present.

3 Methods

We adopt the idea of Dobrynin et al. (2018) to predict the NAO index of the upcoming winter (DJF) based on four predictors, autumn sea ice thickness (SIT), snow depth (SND), sea surface temperature (SST), and stratospheric temperature at 100 hPa (TA100), from the assimilation run of GCFS2.0. The actual values of the predictors are calculated as an area weighted mean of monthly grid cell values, taking into account only grid cells that show a significant correlation to the NAO index. We construct a multiple linear regression estimator for the NAO index that takes all four predictors into account simultaneously.

Multiple linear regression estimators for the three other circulation indices (SCAN, EA/WR, EA) are constructed analogously to the NAO prediction. The literature on driving conditions influencing these indices is rather sparse. However, as already mentioned above, the large scale circulation in the North Atlantic-European sector is a complex interaction of many factors. Boundary fields like the chosen predictors do not impact exclusively on one or another circulation index, but the whole system, exerting a greater or lesser influence on all components. For these reasons, we use the same predictors for SCAN, EW/WR and EA as are proposed for the NAO in Dobrynin et al. (2018).

After having predicted the four circulation indices statistically, in the second step we select the “best” members from the dynamical hindcast ensemble. “Best” is defined here in terms of the Euclidean distance between a dynamical hindcast member’s vector of indices (see subsection 4.1) and our statistically predicted index vector. The “best” members are selected to build a subensemble. The new seasonal hindcasts for SLP, TAS, PR etc. are based only on the subensemble instead of the complete dynamical hindcast ensemble.

3.1 Predictors and Regression

The dynamical seasonal hindcasts for DJF is initialized on November, 1st. We therefore take the October monthly means of SST, SND and TA100 as predictors, as this is the latest information known when the integration starts. For SIT, we use the September monthly mean, because it reflects the annual minimum sea ice extension (Dobrynin et al., 2018).

The correlation between the predictor values and the circulation indices is calculated on grid cell basis. Grid cells, which show a significant positive correlation, are combined to an area weighted sum, as well as grid cells with significant negative correlation. Consequently, each predictor can contribute two exogenous variables to the multi-linear regression. Before entering the regression, the area weighted sums are centered and detrended.

The performance of the proposed estimation procedure is evaluated in subsection 4.2 in the so-called backtesting mode (see Supporting Information), a realistic cross validation setting, where the prediction at a given time is based exclusively on information from its past. In the backtesting mode, we find a high year-to-year variation of the regions, where grid cells with significant correlations between SST and the circulation indices are detected. In some cases this effect leads to a failure in the prediction of the circulation indices. We assume that the relation between SST and the circulation indices is sensitive to the length of the time series, because this effect does not occur when all data is used for the detection. As a remedy, we replace the assimilation time series of SST and SLP (for the calculation of circulation indices) from GCFS2.0 by the respective time series from the latest MiKLip assimilation, which start as early as 1958. The Miklip assimilation is utilized exclusively to detect the significant grid cells. For the calculation of the predictor values we return to the GCFS2.0 assimilation time series.

An ordinary least squares algorithm is performed to estimate the regression coefficients. In order to avoid overfitting, the combination of predictor variables is selected so as to minimize the Mean Squared Error (MSE) of the predicted index in the backtesting mode (see Supporting Information), using a maximum of four predictors.

3.2 Subselection

The subselection of members from the dynamical seasonal hindcast ensemble is based on the statistically predicted circulation indices. To compute the circulation indices realized by each ensemble member, we use the principal components calculated from the assimilation SLP fields. It is very probable that, when applying a PCA to the union of all dynamical seasonal hindcast ensembles, the principal components will not coincide with the GCFS2.0 assimilation. However, for a meaningful comparison between statistically predicted circulation index and its counterpart in a dynamical forecast run, the indices have to refer to the same principal component pattern. We therefore project the dynamical forecast members onto the patterns from the assimilation.

We can now fix the number of circulation indices to be included in the subselection (only one index [NAO], or more than one up to 4). The Euclidean distance is calculated between the index vectors of the dynamical hindcast ensemble members and the vector of statistically predicted indices for a given winter. The Euclidean distance is weighted by the Eigenvalues of the principal components to emphasize the importance of the respective circulation index. Subsequently, the members with the smallest distance to the statistical prediction are selected to build the subensemble. We reiterate the post processing for this subensemble, like generating the ensemble mean, terciles and skill scores for variables of interest like TAS and PR as we have done before on the complete ensemble.

3.3 Selection by Machine Learning Procedures

Further refinements of the subselection that make use of various machine learning procedures are conceivable. We would like to name but a few, details and results of which are described in the Supporting Information. A most obvious refinement would be the weighted mean of the hindcast members according to their proximity to the statistically predicted circulation indices. More sophisticated, a clustering of the vectors of circulation indices would allow for nonlinear interdependencies between the four circulation indices, apart from linear orthogonality imposed by PCA. To improve the achieved stratification of the clusters with respect to TAS (or any other selected parameter), a semi-supervised clustering algorithm or a discriminant analysis could be applied.

4 Results

4.1 Circulation Indices

In this section, we examine our assumption that the seasonal hindcast skill benefits from the inclusion of further circulation indices in the ensemble subselection procedure of Dobrynin et al. (2018). To this end, we repeat their perfect-NAO prediction experiment and compare it to an analogue perfect-circulation indices experiment.

The assimilation run from the current seasonal forecast model GCFS2.0 starts in 1980, hindcasts were provided for start dates in 1990-2017. We consider seasonal means for winter (DJF), such that our time series starts in winter 1980/81 and runs through winter 2017/18, a total of 38 time steps. In order to calculate the winter circulation indices, singular value decomposition is applied to the area-weighted non-standardized anomalies of seasonal SLP over the North Atlantic-European sector ($20-85°$N and $90°$W-$60°$E), Figure 1. Note that the subsequent standardization of the indices does not affect our computations.

Likewise, the ensemble members of the seasonal hindcast ensembles are projected onto the same principal components extracted from the assimilation to calculate the respective circulation indices.

Now, we select those members from the hindcast ensembles, which reproduce the true circulations indices most closely – first only for the NAO index, after that for NAO, SCAN, EA/WR and EA indices. The forecast skill of the full and of the two subensembles is plotted in figure S1 in the Supporting Information.

The improvement in anomaly correlation coefficients (ACC) for SLP in the Euro-Atlantic sector when selecting for all four indices, taken over time at each point separately, is strong. In particular, the zonal band of low predictability between $50°$N and $60°$N, that stands out in the perfect NAO only ensemble, is completely recovered in the four indices ensemble. The ACCs for TAS and PR show considerable improvements, too (Figure S1).

We therefore conclude that the subselection for more than one circulation index is worthwhile–as long as we are able to construct reliable predictors for them.

4.2 Regression

We evaluate the whole estimation and subselection procedure in the backtesting mode, as this is the most realistic setting possible in view of prospective operationalization (see Supporting Information), and the most challenging at the same time. In the following, we will evaluate our predictions and the resulting hindcast skill against the

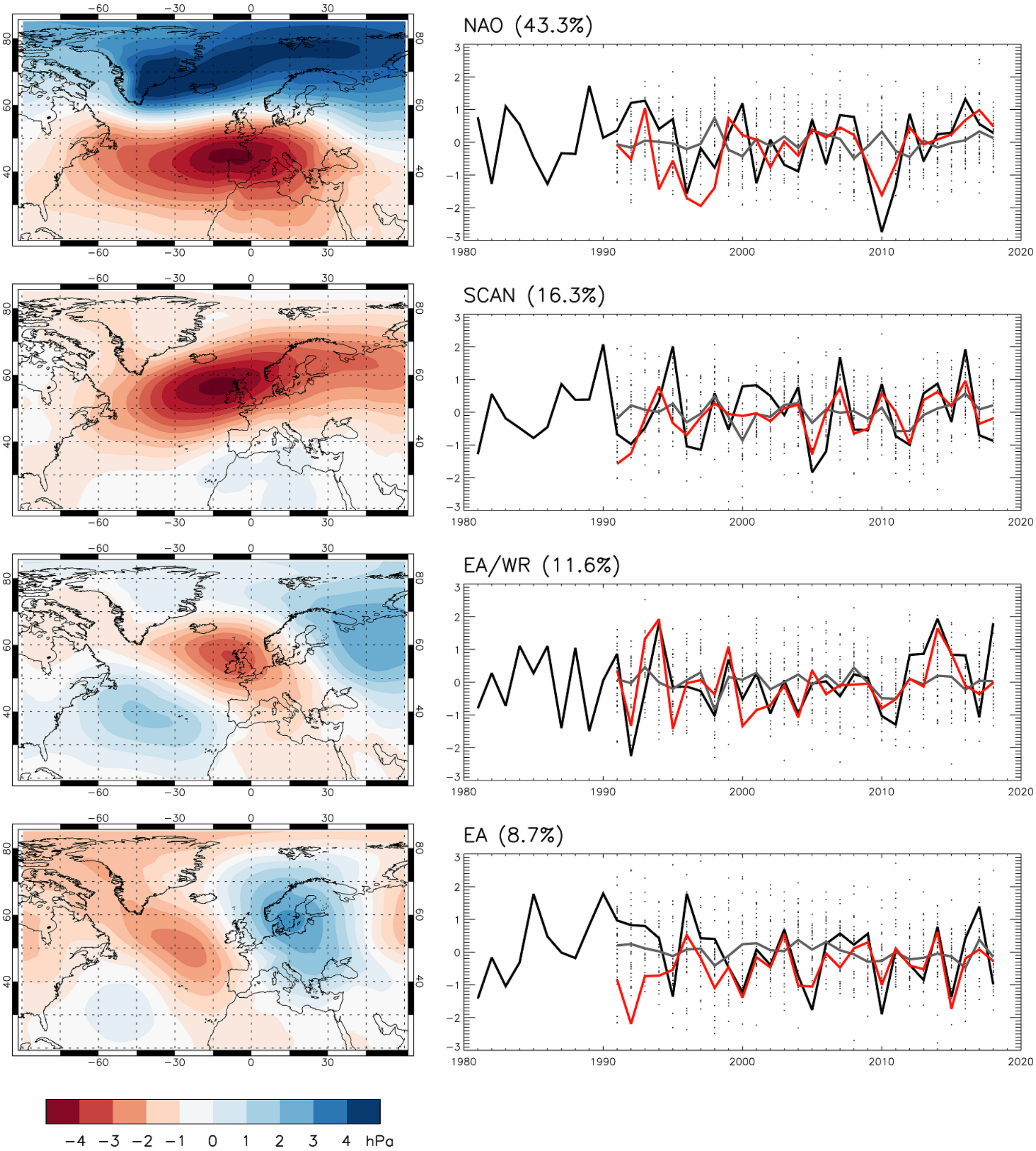

Figure 1. Circulation indices from winters 1980/81-2017/18. Left column: PC loadings for SLP anomalies. Right column: yearly winter PC scores. Black line: GCFS2.0 assimilation, grey line: ensemble mean, grey dots: ensemble members, red line: statistical prediction.

Table 1. Correlation of dynamically hindcasted and statistically predicted to assimilated circulation indices, respectively

	dyn hc	stat pr	SST	SND	SIT	TA100
NAO	0.26/0.15	0.59/0.93	+ −	−	+	
SCAN	0.35/0.56	0.66/0.88	−	−	+ −	
EA/WR	0.35/0.40	0.68/0.73	+ −	+		−
EA	0.23/0.21	0.47/0.80		+ −		+ −

Periods of correlation (DJF 1990/91-2017/18)/(DJF 2003/04-2017/18); Selected predictors: + positively correlated grid cells selected, − negatively correlated grid cells selected

assimilation run of GCFS2.0. We choose the assimilation run over the obvious alternative ERA-Interim for the following reasons: The GCFS assimilation and ERA-Interim are both model assimilations, but the GCFS assimilation was produced with the same model as the hindcasts as opposed to ERA-Interim. The mismatch of the hindcasts will therefore be a priori smaller to the GCFS assimilation, independently of the quality of the hindcasts. Here, we aim to evaluate the relative differences in skill generated by the subselection, so for the moment we set aside model differences between GCFS and ERA-Interim.

The selected predictors and respective correlations between assimilated and statistically predicted circulation indices (as described in subsection 3.1) are listed in Table 1, along with the correlation of the full ensemble mean indices for comparison. Both the algorithm that detects significant predictor grid cells and the least squares estimation are statistical procedures which need a minimum of training data to achieve a certain goodness-of-fit. For early prediction times in the backtest setting, there is only a small amount of data available to train the procedures, which results in poor predictions. We observe that the correlation between the predicted indices and the assimilation strongly depends on the time interval on which the correlation is calculated, with higher values towards the end of the time period. For the purpose of illustration, we give two correlation values for each circulation index in Table 1, one for the winter seasons 1990/91-2017/18, the second for 2003/04-2017/18. A corresponding improvement over time is not apparent in the dynamical ensemble.

In the following we will solely refer to the evaluation period of winters 2003/04-2017/18 to highlight the potential of the proposed procedure. We note that all statistical estimators perform quite well, see Figure 1.

4.3 Subselection

To evaluate whether the subselection leads to an improvement in the seasonal hindcast, we first analyse the anomaly correlation coefficients (ACC) between the ensemble mean of the two hindcasts (subensemble vs. complete ensemble) and the GCFS2.0 assimilation values. Varying the number of selected hindcasts between 4 and 20, we obtained the highest increases in ACC for subensembles of 8 members.

We furthermore varied the number of circulation indices considered in the subselection. It turns out that already the inclusion of the NAO index alone greatly improves the association between hindcast and assimilation (Figure 2). As expected, for the hindcast fields SLP, TAS and PR the ACC increases with each additional circulation index included. The area-weighted average ACC over Europe (10°W-30°E and 35°N-65°N) for SLP is calculated for the full ensemble/NAO-only subselection/4-indices subensemble: 0.24/0.63/0.73. Analogous mean ACCs for TAS amount to 0.41/0.49/0.58 and for PR to 0.22/0.33/0.41.

4.4 Spatial Evaluation of Individual Hindcasts

To further explore the improvement in our temperature hindcasts obtained by subselecting for circulation indices, we compare the individual hindcasts for winter seasons 2008/09, 2009/10 and 2015/16 with the respective GCFS2.0 assimilation in Figure 1. Winters 2009/10 and 2015/16 represent distinctive atmospheric conditions showing unusual values in their circulation indices (2009/10 - very low NAO and low EA, 2015/16 - high NAO and very high SCAN), whereas winter 2008/09 shows average values in all four indices. We find that the assimilation values are poorly reflected in the full ensemble mean indices, except for EA/WR in 2008/09, but they are estimated well by our statistical procedure (Figure 1).

The assimilation temperature anomalies in the three selected winters are quite pronounced. In contrast, the hindcasts anomalies for 2009/10 and 2015/16 from the full ensemble appear quite pale (we concentrate on 10°W-30°E and 35°N-65°N, a region that

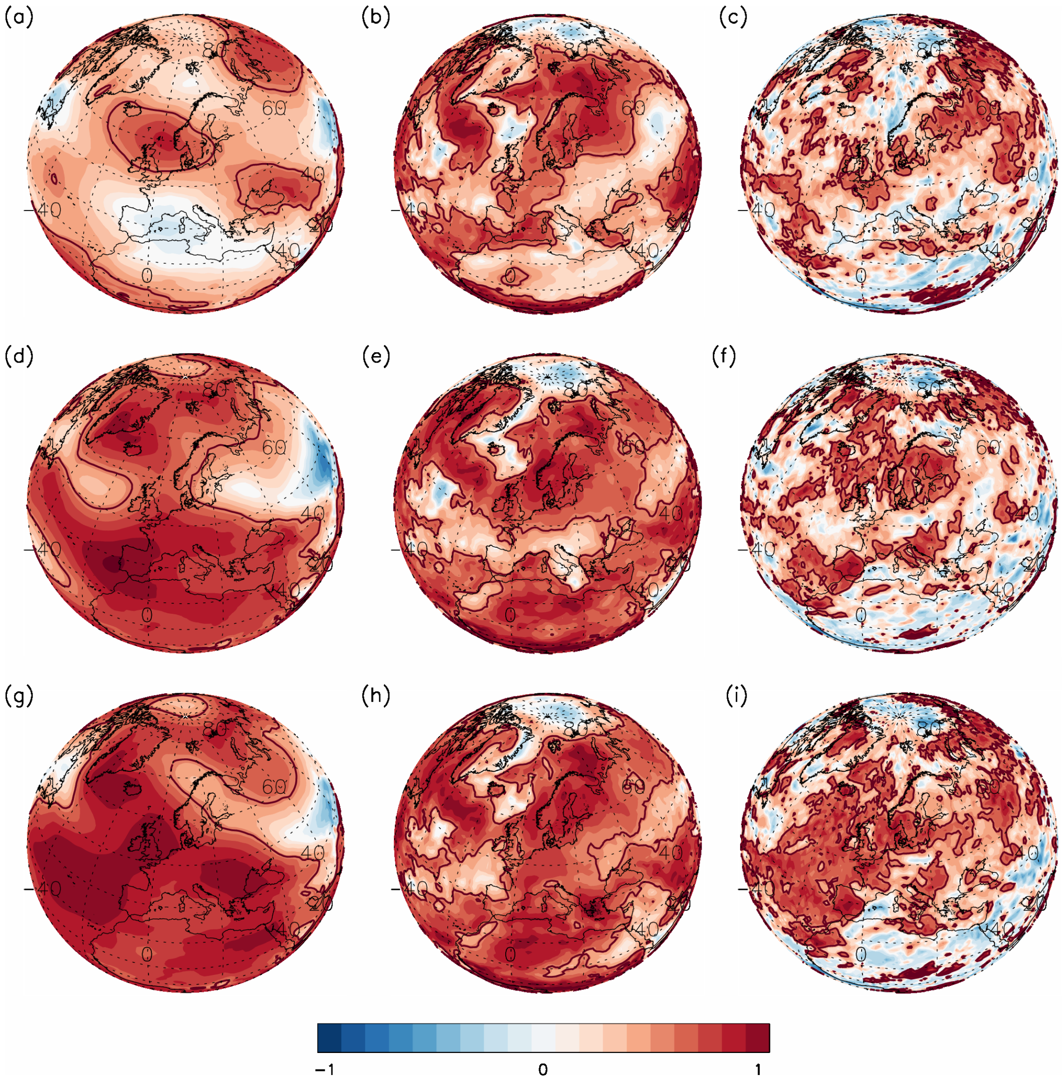

Figure 2. Anomaly correlation coefficients between hindcast ensemble means and assimilation for winters 2003/04-2017/18. 1st row: complete ensemble, 2nd row: subselection for NAO, 3rd row: subselection for NAO, SCAN, EA/WR, EA. Left column: SLP, center column: TAS, right column: PR. Regions, where the ACC is significantly positive to the 95% level (critical value 0.441), are contoured in dark red.

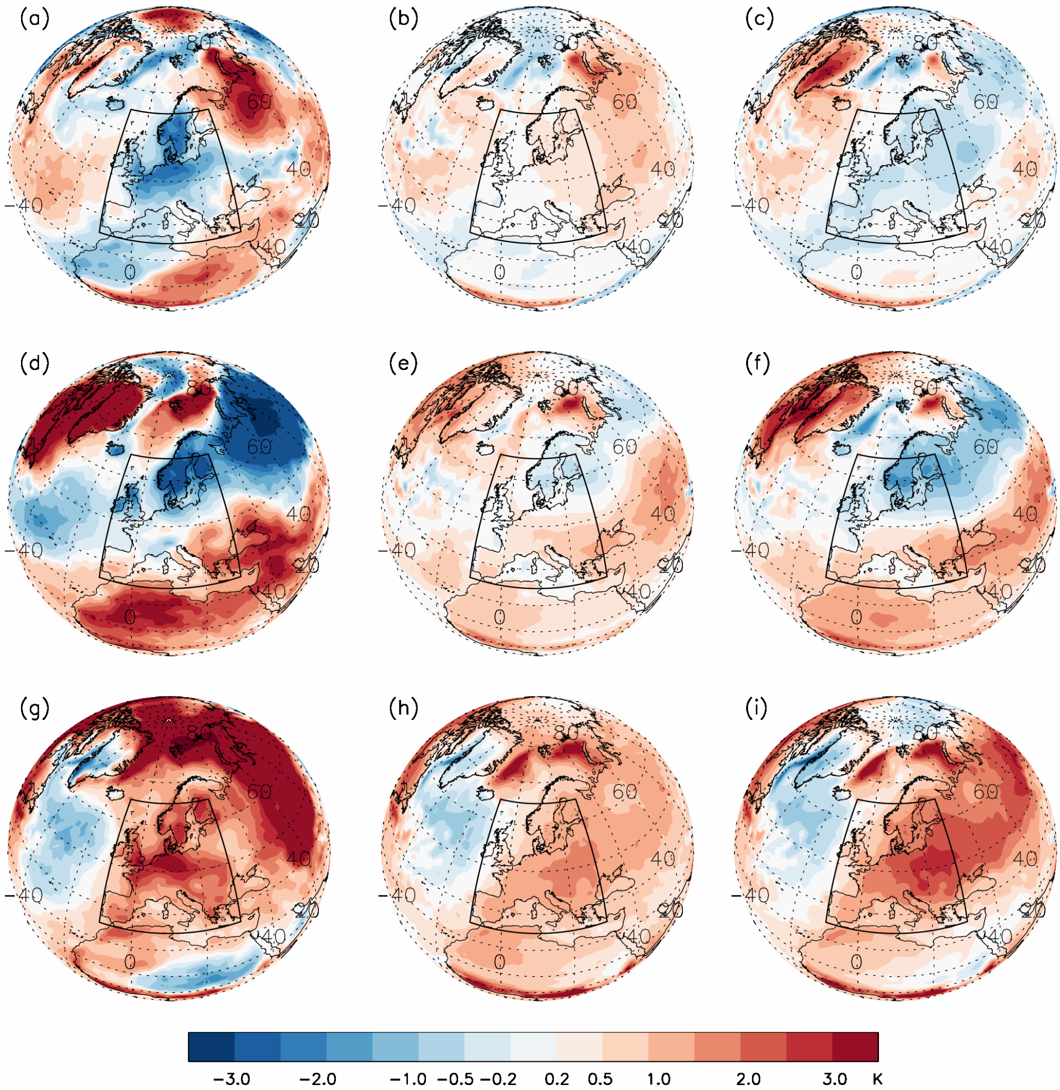

Figure 3. Temperature anomalies for 2008/09 (1st row), 2009/10 (2nd row) and 2015/16 (3rd row). Left column: GCFS2.0 assimilation, center column: full ensemble, right column: subensemble. Black contoured rectangle: the target area 10°W-30°E, 35°N-65°N

constitutes a natural target for the German Meteorological Service, see Figure 3). For 2008/09 the full ensemble mean hindcast fails completely to capture the generalized cold anomaly. After subselection, in 2009/10 the spatial pattern of anomalies is very well reproduced and also the warm hindcast anomalies for 2015/16 are increased and much closer to the analysed ones. For winter 2008/09, the subselected forecast shows a cold anomaly reversing the full ensemble hindcast. However, all subselected anomalies are still weakly pronounced in amplitude comparing to the assimilation run (Figure 3).

To quantify the goodness-of-fit of the individual full and subselected ensemble hindcasts, we evaluate the Structural Similarity Index (SSIM) (see Supporting Information and Z. Wang et al. (2004)) over the target region (Figure 3). Within this region, we weight grid cell contributions to the SSIM by area. As might be suspected from visual inspection, SSIM between TAS hindcasts and assimilations is markedly increased by subsampling. A further improvement is obtained by simple rescaling, which results in an amplification of both the cold and warm anomalies towards more realistic values, opening prospects for more sophisticated bias correction methods (see Table S1 in the Supporting Information).

Although the SSIM increase by subselection with regard to TAS is most pronounced in the selected years, the average skill for TAS SSIM in 2004-2018 has also more than doubled (Supporting Information Table S2). For SLP and PR the increase obtained by subselection is even more and slightly less pronounced, respectively. The results obtained using other selection procedures (subsection 3.3), which partly surpass the improvements of the simple subselection by far, are listed in the Supporting Information (Table S2).

5 Summery and Discussion

We have constructed an ensemble selection procedure based on the statistical prediction of the four leading principal components of SLP in the North Atlantic-European sector, which leads to a substantial improvement of seasonal hindcast skill for winter (DJF) hindcasts of SLP, TAS and PR compared to the full ensemble mean hindcasts. This method is evaluated in the backtesting mode, with average anomaly correlation over Europe for SLP, TAS and PR of 0.73, 0.58 and 0.41, respectively. The statistical predictions rely solely on the autumn states of four drivers of atmospheric circulation, which are known at the time the dynamical model integration starts. The procedure is therefore fully applicable to operational forecasts.

The presented subsampling method is tailored to improve the seasonal hindcasts in winter over Europe, only. Skill over other regions and seasons is thus possibly degraded. Nonetheless, an analogue approach aiming at other regions and seasons is conceivable.

We have to assume that the relationships between the predictors, the circulation indices and the seasonal climate that we exploit in our subselection might be subject to climate variability as well as climate change. In the long run, strategies accounting for such non-stationarity have to be developed.

Acknowledgments

The authors would like to thank Klaus Pankatz for providing the MiKlip assimilation data and the corresponding description, and Johanna Baehr for valuable scientific discussions and comments on the manuscript. Hindcast data used for this study are available under https://cds.climate.copernicus.eu/cdsapp#!/dataset/seasonal-monthly-single-levels?tab=form.

References

Baehr, J., Fröhlich, K., Botzet, M., Domeisen, D. I. V., Kornblueh, L., Notz, D., ... Müller, W. A. (2015, 5). The prediction of surface temperature in the new seasonal prediction system based on the mpi-esm coupled climate model. *Climate Dynamics*, *44*(9-10), 2723-2735. doi: 10.1007/s00382-014-2399-7

Baker, L., Shaffrey, L., & Scaife, A. (2018). Improved seasonal prediction of uk regional precipitation using atmospheric circulation. *International Journal of Climatology*, *38*, e437–e453.

Baker, L., Shaffrey, L., Sutton, R., Weisheimer, A., & Scaife, A. (2018). An intercomparison of skill and overconfidence/underconfidence of the wintertime north atlantic oscillation in multimodel seasonal forecasts. *Geophysical Research Letters*, *45*(15), 7808–7817.

Barnston, A., & Livezey, R. (1987). Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Monthly Weather Review*, *115*(6), 1083–1126.

Cohen, J., Coumou, D., Hwang, J., Mackey, L., Orenstein, P., Totz, S., & Tziperman, E. (2019). S2s reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. *Wiley Interdisciplinary Reviews: Climate Change*, *10*(2), e00567.

Comas-Bru, L., & McDermott, F. (2014). Impacts of the ea and sca patterns on the european twentieth century nao-winter climate relationship. *Quarterly Journal of the Royal Meteorological Society*, *140*(679), 354–363.

Dobrynin, M., Domeisen, D., Müller, W., Pohlmann, H., & Baehr, J. . (2018). Improved teleconnection-based dynamical seasonal predictions of boreal winter. *Geophysical Research Letters*, *45*(8), 3605–3614.

Hall, R. J., Scaife, A. A., Hanna, E., Jones, J. M., & Erdélyi, R. (2017). Simple statistical probabilistic forecasts of the winter nao. *Weather and Forecasting*, *32*(4), 1585-–1601.

Hurrell, J., Kushnir, Y., Ottersen, G., & Visbeck, M. E. (2003). *The north atlantic oscillation: Climatic significance and environmental impact.* Washington, DC: American Geophysical Union.

Hurrell, J. W. (1995). Decadal trends in the north atlantic oscillation: Regional temperatures and precipitation. *Science*, *269*(5224), 676–679.

Iglesias, I., Lorenzo, M., & Taboada, J. (2014). Seasonal predictability of the east atlantic pattern from sea surface temperatures. *PLoS ONE*, *9*(1), e86439.

Kretschmer, M., Coumou, D., Donges, J., & Runge, J. (2016). Using causal effect networks to analyze different arctic drivers of midlatitude winter circulation. *Journal of Climate*, *29*(11), 4069-4081.

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., ... Roeckner, E. (2018). Developments in the mpi-m earth system model version 1.2 (mpi-esm1.2) and its response to increasing co_2. *J. Adv. Modelling Earth Syst. (JAMES)*, *11*, 998–1038. doi: 10.1029/2018MS001400

Müller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., ... Marotzke, J. (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*(7), 1383-1413. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2017MS001217` doi: 10.1029/2017MS001217

Ossó, A., Sutton, R., Shaffrey, L., & Dong, B. (2018). Observational evidence of european summer weather patterns predictable from spring. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(1), 59-63.

Pohlmann, H., Müller, W. A., Bittner, M., Hettrich, S., Modali, K., Pankatz, K., & Marotzke, J. (2019). Realistic quasi-biennial oscillation variability in historical and decadal hindcast simulations using cmip6 forcing. *Geophys. Res. Lett.*, *46*, 14118-14125.

Rust, H., Richling, A., Bissolli, P., & Ulbrich, U. (2015). Linking teleconnection patterns to european temperature - a multiple linear regression model. *Meteorologische Zeitschrift*, *24*(4), 411–423.

Thompson, D., Lee, S., & Baldwin, M. (2003). *Atmospheric processes governing the northern hemisphere annular mode/north atlantic oscillation* (Vol. 134).

Vihma, T., Graversen, R., Chen, L., Handorf, D., Skific, N., Francis, J., ... Overland, J. (2018). Effects of the tropospheric large-scale circulation on european winter temperatures during the period of amplified arctic warming. *International Journal of Climatology*, *40*(1), 509-529.

Wang, L., Ting, M., & Kushner, P. (2017). A robust empirical seasonal prediction of winter nao and surface climate. *Scientific Reports*, *7*(279).

Wang, Z., Bovik, A., Sheikh, H., & Simoncelli, E. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600-612.

Supporting Information for "Seasonal Forecasts of Winter Temperature Improved by Higher-Order Modes of Mean Sea Level Pressure Variability in the North Atlantic Sector"

Clementine Dalelane, Mikhail Dobrynin, Kristina Fröhlich

Deutscher Wetterdienst (DWD)

Contents of this file

1. Backtesting
2. Stuctural Similarity Index
3. Selection by Machine Learning Procedures
4. Table S1: SSIM of hindcasted to assimilation anomalies in individual years
5. Table S2: Average SSIM of hindcasted to assimilation anomalies in 2004-2018
6. Figure S1: Anomaly correlation coefficients for perfect subselection

Backtesting

Backtesting is a causal type of cross-validation, in which the forecasting procedure is applied only to data prior to the forecast time. In order to reforecast a circulation index at time t, we perform the whole procedure of selection of significant predictors and least squares estimation on the data for $s = 1 \ldots t-1$.

We note that the Principal Component Analysis is not included in the backtesting mode. That is because PCA is essentially a transformation of coordinates with the objective to aggregate as much variance as possible into a small number of directions. A unique "true" set of principal components does not exist. Furthermore the estimation procedure is not affected by small changes in the principal components as long as the correlation between the predictors and the indices is preserved. So we use fixed circulation indices calculated from the whole time series of SLP fields to investigate the properties of the proposed subselection.

Structural Similarity Index

To summarize the performance of the hindcasts, we introduce the Structural Similarity Index (SSIM), a concept developed in the context of image processing (Wang et al., 2004). The SSIM is used to measure the similarity between two images, in our case the similarity between the hindcast and assimilation fields. It combines three important aspects of spatial goodness-of-fit, which in climatological forecast validation are usually measured and assessed separately: mean, variance and correlation.

Let x and y be the two fields to compare, c_1 and c_2 small constants. Then

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

where μ_x, μ_y are the respective spatial means, σ_x^2, σ_y^2 the spatial variances and σ_{xy} is the spatial covariance.

SSIM satisfies the non-negativity, identity of indiscernibles, and symmetry properties. The resultant SSIM index is a decimal value between -1 and 1, and value 1 is only reached for two identical fields and therefore indicates perfect structural similarity. A value of 0 indicates no structural similarity.

As the SSIM includes terms of mean and variance, it is improved by linear bias-adjustment (rescaling), although this does not alter the ACC.

SSIM values of the full, subselected and rescaled-subselected ensemble hindcasts corrsponding to the winter seasons 2008/09, 2009/10 and 2015/16 for the variables TAS, SLP and PR, calculated over certain regions, are listed in Table S1. As the smoothness and spatial correlation of these climatological parameters are very different, we choose larger/smaller regions for the spatial average, which are nevertheless all centered over

Germany as it constitutes the natural target for the German Meteorological Service. For TAS this is the region between 10°W-30°E and 35°N-65°N, for SLP 50°W-47°E and 23°N-85°N, and for PR 6°-16.5°E and 46.75°N-56°N.

References

Wang, Z., Bovik, A., Sheikh, H., & Simoncelli, E. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600-612.

Selection by Machine Learning Procedures

A *weighted subselection* is realized by the application of a radial Epanechnikov-kernel to the (Eigenvalue-weighted) Euclidian distances between statistically predicted and dynamically hindcasted circulation-index vectors x_s and x_d. Let W be the diagonal matrix of Eigenvalues of the circulation indices obtained from the principal component analysis, then the Epanechnikov-kernel with bandwidth h is defined as

$$K_h(x_s, x_d) = \frac{3}{4h}\left(1 - \frac{\|x_s - x_d\|_{2W}^2}{h^2}\right), \text{ where } \|x_s - x_d\|_{2W}^2 = x_s^T W^{1/2} x_d$$

The weighted subensemble is realized by the weighted sum of all hindcast ensemble members. Best results where obtained with three circulation indices and a bandwidth of $h = 87$ (see Table S2).

The *clustering of circulation indices* allows for nonlinear interdependencies between the four circulation indices, apart from linear orthogonality imposed by PCA.

To obtain an unsupervised classification of the vectors of circulation indices, a K-means algorithm (with Egenvalue-weighted Eucliadian distance) is applied to the circultion-index vectors $x_a^1 ... x_a^T$ for all winters from the assimilation run. The algorithm is initialized by intermediate index values. Subsequently, the statistically predicted and the dynamically hindcasted vectors for a specified winter are assigned to their nearest cluster. The subensemble is then composed of those hindcast members that pertain to the cluster indicated by the statistical prediction. The resulting improvements w.r.t SSIM(TAS), achieved for three index vectors and five clusters, are listed in table S2.

Algorithm of clustered selection

1. The circulation-index vectors from assimilation $x_a^1...x_a^T$ are clustered, clusters $C_1...C_K$ are obtained

2. The statistically predicted index vector x_s for the specified winter is assigned to the nearest cluster C_k – this is the staistically predicted cluster

3. The hindcasted index vectors $x_d^1...x_d^{30}$ for the specified winter are assigned each to their nearest cluster

4. The subensemble is composed of those hindcast members that fall into the staistically predicted cluster C_k

To improve the stratification of the clusters w.r.t. some target variable (TAS in our case), in *semi-supervised clustering* the training sample is augmented by the values of the target variable assumed in the training sample $y_a^1...y_a^T$, such that $\tilde{x}_a^t = (x_a^t, y_a^t)$. The clustering procedure is otherwise identical to the unsupervised clustering. In the classification of a statistical prediction x_s^t, the target value is of course unknown and the assignment is based on the circulation indices only. In our case, we generate the target variables from the TAS field by principal component analysis. The resulting scores of the leading TAS PCs are introduced as target variables into the K-means algorithm, in addition to the circulation indices resulting from the PCA of the SLP fields. The hindcasts are processed analogously to the un-supervised clustering (see results for the best parameter combination [4 circulation indices, 2 TAS PCs, 4 clusters] in Table S2).

Linear distcriminant analysis is a supervised classification procedure that optimally seperates two or more classes of objects on the basis of observable variables. The classification of the training sample has to be known in advance. We define three classes w.r.t. the same two TAS PCs used above. The algorithm finds the linear partition in the space of circulation indices that best predicts the given classification. According to the statistically predicted circulation indices, a class is selected along with the corresponding hindcasts. Best discrimination results where obtained using all 4 circulation indices (Table S2).

All selection results in Table S2 have been generated in the Backtesting mode.

Table S1. SSIM of hindcasted to assimilation anomalies in selected winters of full/subselected/subselected+rescaled ensemble

	DJF 2008/09	DJF 2009/10	DJF 2015/16
TAS	0.01/0.24/0.30	0.27/0.75/0.82	0.42/0.62/0.70
SLP	-0.09/0.51/0.54	0.08/0.32/0.39	0.25/0.61/0.68
PR	0.01/0.25/0.24	0.30/0.35/0.37	0.18/0.54/0.69

Table S2. Average SSIM of hindcasted to assimilation anomalies during winters from 2003/04 through 2017/18

	full	Sub	SubR	SubWR	ClAnaR	sClAnaR	DisAnaR
TAS	0.12	0.29	0.31	0.36	0.34	0.33	0.38
SLP	0.11	0.34	0.37	0.45	0.40	0.33	0.41
PR	0.06	0.10	0.10	0.07	0.17	0.12	0.12

Selection types: full-full ensemble, Sub-subselection of 8 best members, SubR-rescaled subselection of 8 best members, SubWR-rescaled weighted subselection, ClAnaR-rescaled subselection according to unsupervised cluster analysis, sClAnaR-rescaled subselection according to semi-supervised cluster analysis, DisAnaR-rescaled subselection according to discriminant analysis

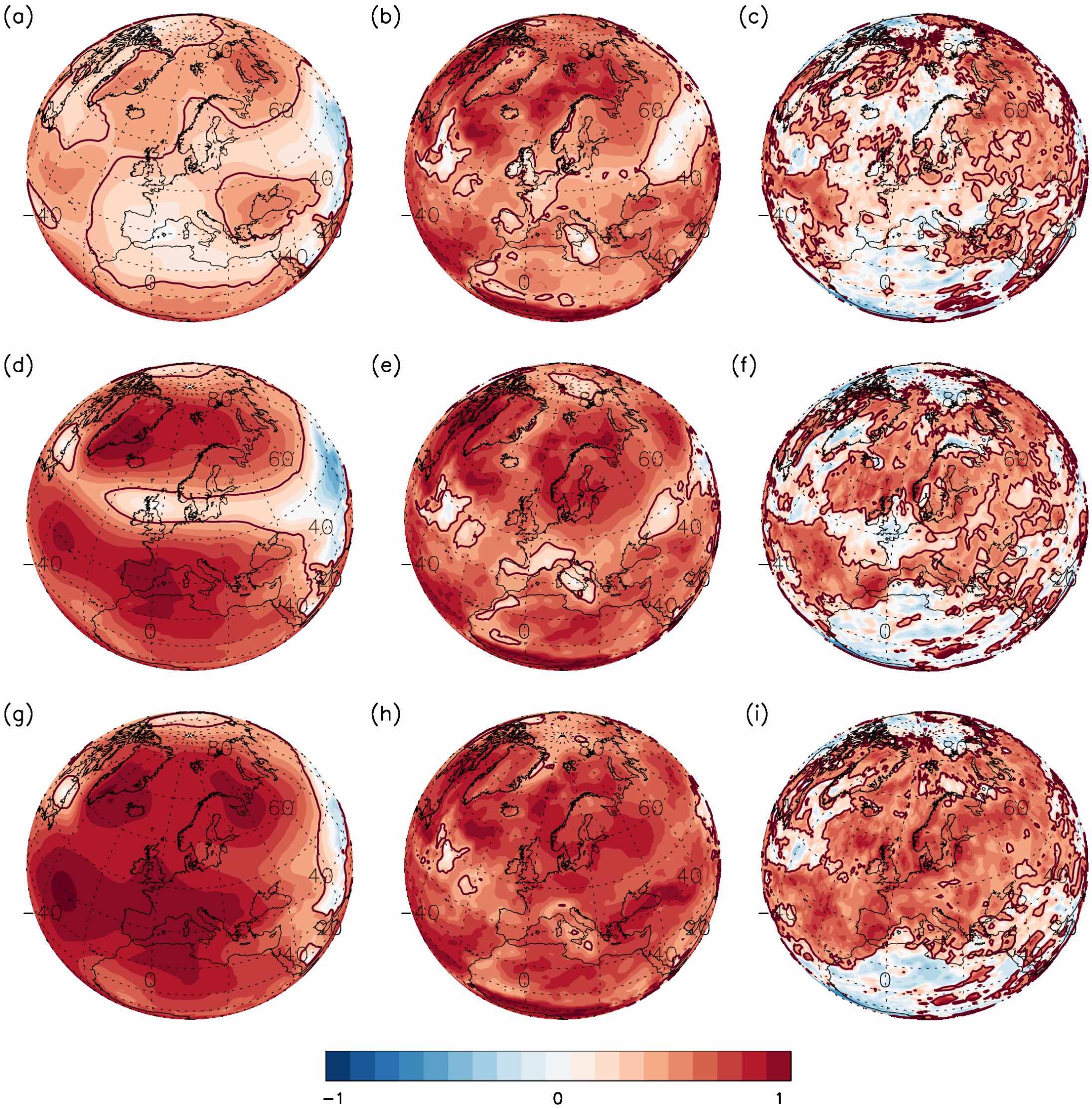

Figure S1. Anomaly correlation coefficients between ensemble means and assimilation. 1st row: complete ensemble, 2nd row: subselection for perfect NAO, 3rd row: subselection for perfect NAO, SCAN, EA/WR and EA. Left column: SLP, center column: TAS, right column: PR. Regions, where the ACC is significantly positive to the 95% level (critical value 0.271), are contoured in dark red.