Statistical learning methods as a basis for skillful seasonal temperature forecasts in Europe

Matti Kämäräinen[1], Petteri Uotila[3], Alexey Yu. Karpechko[2], Otto Hyvärinen[1], Ilari Lehtonen[1], Jouni Räisänen[3]

Corresponding author: Matti Kämäräinen[1] (matti.kamarainen@fmi.fi)

[1]Weather and Climate Change Impact Research, Finnish Meteorological Institute, Helsinki, Finland

[2]Meteorological Research, Finnish Meteorological Institute, Helsinki, Finland

[3]Institute for Atmospheric and Earth System Research, University of Helsinki, Helsinki, Finland

Keywords: seasonal forecasting, statistical learning, predictability, skill, lagged predictors, ensemble prediction, deterministic forecast, LASSO regression

Abstract

A statistical learning approach to produce seasonal temperature forecasts in western Europe and Scandinavia was implemented and tested. Leading principal components (PCs) of sea surface temperature (SST) and geopotential at the 150 hPa level (GPT) were derived from reanalysis datasets and used at different lags (1 to 5 seasons) as predictors. Random sampling of both fitting years and potential predictors together with Least Absolute Shrinkage and Selection Operator regression (LASSO) was used to create a large ensemble of statistical models. Applying the models to independent test years shows that the ensemble performs well over the target areas, and that the ensemble mean is more accurate than the best individual ensemble member on average. Especially skillful results were found for summer and fall, anomaly correlation coefficient values ranging between 0.41 and 0.68 for these seasons. Correct simulation of decadal trends, using long enough time series for fitting (70 years), and using lagged predictors increase the prediction skill. Decadal scale variability of SST, most importantly the Atlantic multidecadal oscillation (AMO), as well as different PCs of GPT are the most important individual predictors among all predictors. Both SST and GPT bring equally much predictive power, albeit their importance is different in different seasons.

1. Introduction

Skillful seasonal forecasts can help various weather sensitive sectors to anticipate for weather related risks (Clark et al. 2017; Tauser and Cajka 2014; De Cian et al. 2013). For that reason, the predictability of temperature, precipitation, and, for example, the North Atlantic Oscillation index (NAO) in monthly and seasonal time scales has been an active research topic in Europe. Several potential sources of predictability have been reported, including sea surface temperatures (SSTs) in various basins (e.g., Kolstad and Årthun 2018; Smith et al. 2016; Toniazzo and Scaife, 2006; Brönnimann, 2007; Rodwell and Folland, 2002), continental snow cover (Allen and Zender 2011; Cohen and Jones 2011; Cohen and Entekhabi 1999), stratospheric geopotential (GPT) or winds (Jia et al. 2017; Wang et al. 2017; Scaife et al. 2016; Brönnimann et al. 2016), sea ice cover (Liptak and Strong 2014; Vihma et al. 2014), and soil moisture (Orth and Seneviratne 2014; van den Hurk et al. 2012). These parameters share two important properties: they vary slowly in time, and they can act as forcings for the troposphere. Both properties increase the predictive power of seasonal forecasting models.

Statistical modeling is an appealing approach to produce seasonal forecasts, because it is computationally cheap to run compared to numerical-dynamical models, which require massive hardware for parallelization, long running times, and large storage capacity for storing the input and output data. In addition to that, even though skillful in predicting the wintertime NAO (Baker et al. 2018b), current operational dynamical models do not always perform particularly well in simulation of fundamental climate variables, such as temperature and precipitation over Europe (Mishra et al. 2018; Weisheimer and Palmer 2014). Assuming that predictability is an inherent part of the climate system in seasonal time scales, and that dynamical models are only partly capable of utilizing it, investigations of better statistical approaches remain useful. Furthermore, uncovered statistical linkages may provide valuable insights into physical behaviour of the climate system.

Previously, both dynamical and statistical models have been used to predict, for example, the precipitation over British Isles (Baker et al. 2018a; Ossó et al. 2017), the precipitation in Europe (Dunstone et al. 2018; Totz et al. 2017), the wintertime NAO index (Dobrynin et al. 2018; Hall et al. 2017; Scaife et al. 2014; Stockdale et al. 2015), snow accumulation over the Alps (Förster et al. 2018), the sea ice cover in the Baltic Sea (Karpechko et al. 2015), wintertime European temperatures (Folland et al. 2012), and monthly temperatures at various mid-latitude locations (Karpechko 2015).

Machine learning and statistical learning methods are becoming increasingly popular in different applications of atmospheric sciences (e.g., Sprenger et al. 2017; Ukkonen et al. 2017). However, they have not been used very comprehensively in statistical seasonal forecasting yet, probably because the most advanced deep learning methods require vast amounts of training data to learn the nonlinear relationships between predictors and predictands. In the framework of seasonal forecasting, the typical available number of years (i.e., the number of samples n), is only less than one hundred. This small n may effectively disable fitting of deep learning models, such as artificial neural networks. However, statistical learning methods vary widely, and some methods are more suitable to smaller datasets than others. This paper is a step towards utilizing the potential and power of statistical learning, using as robust statistical learning methods as possible to avoid overfitting.

The model consists of an ensemble of individual regression models created with random sampling and regularization of predictors. Principal components (PCs) of large-scale predictor parameters are used as potential predictors to utilize teleconnections from global or hemispheric domains. The regularization, or shrinkage, is based on the Least Absolute Shrinkage and Selection Operator regression (LASSO; e.g., Hastie et al. 2009). The skill of the proposed procedure in producing seasonal and areal mean temperature forecasts is evaluated over a set of independent test years in different target domains (Figure 1).

The paper is organized as follows. In sections 2a–c the predictor and predictand variables are presented, followed by description of the statistical methods in sections 2d–g. In section 3 the skill of the ensemble is evaluated over target domains and seasons (3a), the importance of details in modeling is shown (3b), and automatically selected predictors are analysed (3c). Discussions and conclusions are collected in sections 4 and 5.

2. Materials and methods

The shortness of time series is a major challenge in statistical seasonal forecasting. It is tempting to use the years of the satellite-era, often considered to begin in late 70s, in training the statistical models because the quality of the reanalysis datasets increases when satellite observations are included. However, this period covers only ~40 years, which might be too short to obtain statistically significant results, especially when seasonal or monthly averaging is used in data aggregation, such that each month or season is represented only 40 times in the data. Potentially important, decadal-scale variations of the climate system can not be identified from such short a sample. Isolating separate test years from the original set of years makes the training sample even smaller. Finally, the chaotic, unpredictable part of the variability is often large compared to the predictable part, which makes fitting of statistical models uncertain when the signal-to-noise ratio is too small.

Moreover, the selection of the test years for validation of the model is not trivial. First, to be able to reliably measure the performance statistics, the number of samples (i.e., years in this case) should not be too small as the range of uncertainty estimates becomes impractically large. Second, to avoid the potentially existing autocorrelations in the predictors and predictands affecting the results too much, a contiguous and probably even isolated period should be used, with some buffer years between the fitting and testing years. Third, the most recent years should be selected as a test sample, so that the potential changes in predictability, caused, for example, by the ongoing climate change on centennial scale, or decadal scale ocean processes, could be seen and generalized to represent the current state of the climate system as well as possible. In other words, an adequately *long*, *contiguous*, and *representative* period of time is required for testing (Tashman 2000). Sometimes different cross-validation procedures have been used as an alternative for independent test years, but overfitting can still happen if the same years are used in cross-validation and predictor selection, and as a result, the application of those models to truly independent years fails (e.g., DelSole and Shukla 2009).

The ERA-20C reanalysis dataset (Poli et al. 2016) was used as the source for predictor data in this study. As an alternative and complementary input data, the version 2C of the Twentieth Century Reanalysis (20CRv2c; Compo et al. 2011) was used. All results shown are based on ERA-20C unless otherwise stated. The data were downloaded in monthly time resolution and in their original spatial resolutions (ERA-20C: 1.125°×1.125°; 20CRv2c: 2°×2°). Leading principal components of global SST and Northern Hemispheric GPT at the 150 hPa level were extracted from both reanalyses to be used as potential predictors. The ERA-Interim reanalysis data (Dee et al. 2011) was used to validate the GPT data of other reanalyses. Near-surface temperatures from the observational HadCRUT4 dataset (5°×5°; Morice et al. 2012) were averaged over target domains and seasons to be used as predictands. Details of the treatment of the parameters are explained in the next sections.

In seasonal forecasting all predictors in the predictor matrix, consisting of p potential predictors of length n, are more or less weak in explaining the predictand variables. For this reason it is important to pay attention not only to the choice of the statistical method, but also to the dimensions of the predictor matrix. To minimize the risk of overfitting, it is necessary to use as long time series as possible (a large number of samples n) for fitting. As demonstrated by Hastie et al. (2009), risk of overfitting increases rapidly when $p \approx n$ or $p > n$ (high dimensional dataset), but the risk is smaller when $p < n$ (low dimensional dataset). As n increases, it is possible to increase also the number of potential predictors p and still have $p < n$. On the other hand, using too uncertain data for fitting yields poorly fitted models. The uncertainty of reanalyses is the largest in the earliest years because of sparseness of the observations used in the assimilation (Poli et al. 2016; Compo et al. 2011), and for that reason the years before 1915 were excluded from the centennial scale reanalyses. Excluding longer periods increases the ratio $\frac{p}{n}$ (i.e., the dimensionality of the predictor matrix) too much, and for that reason it was found to decrease the accuracy of the model, as shown in section 3b.

a. Selection of predictor parameters

As mentioned in Introduction (chapter 1), different observational SST data have traditionally been used in statistical seasonal forecasting, and their predictive power is well known. Long time series are easily accessible either as derived indices describing SST variability in different locations, or as global gridded products. For these reasons SST is a natural choice to be used in this study also.

Additionally, variables describing stratospheric circulation are becoming popular, because they contain exploitable signals in the seasonal scale, and because their accessibility is good nowadays, as modern reanalyses contain also pressure level information in addition to the surface layer data. Only surface data have been used in the assimilation of the ERA-20C and 20CRv2c reanalyses, and because upper-air observations were not assimilated, there is a potential risk that the derived stratospheric circulation would be too uncertain to be used as a predictor parameter for seasonal forecasting. Earlier, Gerber and Martineau (2018) validated the annular modes of the atmospheric circulation in several reanalysis datasets, including ERA-20C and 20CRv2c, and found their representation of the modes to be quite accurate, especially in the Northern Hemisphere. To further validate the seasonal 150 hPa level GPT data for this study, the ERA-Interim reanalysis geopotential anomalies for that level were compared to the GPT anomalies of ERA-20C and 20CRv2c in 1979–2010. The assimilation model of ERA-Interim includes also the upper-air observations, and for that reason it can be assumed to be reliable at different altitudes. High consistency was found in the representation of GPT between different reanalyses, as seen in Figure 2. For example, the temporal correlation between ERA-20C and ERA-Interim GPT anomalies is 0.80 at its minimum, and 0.93 on average. Based on this analysis the 150 hPa GPT is accurately represented in reanalysis datasets at least for the recent decades, and is therefore potentially useful for training the seasonal forecasting models.

In addition to GPT and SST, other potential predictor parameters, such as sea ice and snow cover have been used in seasonal forecasting previously. See Appendix A for the reasons why they are not included in this study.

b. Extraction of predictor variables

Both parameters, SST and GPT, were aggregated to seasonal time resolution using temporal averaging over the standard 3-monthly DJF, MAM, JJA, and SON seasons. Seasonal anomalies were then calculated by subtracting the 1915–2010 seasonal means from each grid cell separately for both parameters, followed by application of latitude weighting of the data, using $w(\phi) = \sqrt{cos(\phi)}$, to decrease the influence of the high-latitude grid cells in the results (Wilks 2011, section 12.2). Finally, to remove the centennial scale climate change signal, the 1915–2010 linear trends were subtracted from the predictor data[1], separately for

[1] Alternatives for that procedure exist in literature. To implement, for example, the low-pass filtering approach of Huang et al. (1996), the local mean temperature for the recent past (e.g. past 10 years) should be added as an

each season.

After the aforementioned preprocessing steps, PC time series with $N_{PCs} \in$ [1, 5] for SST and with $N_{PCs} \in$ [1, 3] for GPT were calculated from the seasonal anomaly fields of each parameter using year-round data and global domain for SST and Northern Hemispheric domain for GPT (Figures 3 and 4). Altogether 37% of the total variance of the SST data is explained by the leading five PCs. The distribution of the explained variance proportion in the GPT data is different: 56% of the variance was explained already by the first three components.

The SST PC time series were then lagged with $N_{lags} \in$ [1, 5] to be able to take into account different, possibly prolonged causal relations of the predictors. In 3-monthly data the N_{lags} value of 2, for example, means that predictor values from two seasons ago are used to forecast the predictand values of the target season. A total of 5 components × 5 lagged steps = 25 potential PC predictors were derived from the SST parameter. Changes in stratospheric circulation, including, for example, sudden stratospheric warming events (SSWs; e.g., Kidston et al. 2015), bring exploitable signals for the GPT parameter, and they typically affect the troposphere no longer than for two seasons. For this reason the number of lags was reduced to $N_{lags} \in$ [1, 2] for GPT. The total number of potential GPT predictors was then 3 components × 2 lagged steps = 6, and altogether 25 + 6 = 31 potential PC predictors were

additional predictor in the statistical model. However, defining the optimal averaging period objectively is not trivial, and additionally, this approach would hamper the usefulness of the potentially valuable decadal scale signals in the data.

extracted from the SST and GPT parameters. Finally, all potential predictors were scaled to N(0,1) (i.e., to have zero mean and unit variance), because LASSO regression can not be applied without standardization of the predictors.

The first SST component, SST1, describes the El Niño–Southern Oscillation (ENSO) index. The second component, SST2, represents mostly the Atlantic Multidecadal Oscillation (AMO). The third and fourth are both connected to the Pacific Decadal Oscillation (PDO) and to the North Pacific Gyre Oscillation (NPGO) indices. See Appendix B for the explanation how different PCs of SST were linked to the previously identified SST variability.

c. Aggregation of the predictand variables inside target domains

Temporal and spatial aggregation inside the study domains (Figure 1) was applied to the monthly HadCRUT4 temperature. The data were first season-averaged over the seasons. After that, the seasonal cycle was removed by subtraction of the seasonal means, and detrending was applied by removing the linear 1915–2010 trend from each season separately. Finally, the areal means were calculated to form one predictand time series per domain.

To study the representativeness of the areal mean temperature inside the target domains, the correlation coefficient between the areal mean temperature anomaly and each individual grid cell anomaly in SC and WE were calculated. Mean correlation over all cells was 0.88 in SC and 0.78 in WE, and the weakest correlations, > 0.58, were found in the offshore grid cells. Based on this analysis, the areal mean temperature represents well the temperature variability over land areas in the smaller SC and WE domains, and the mean temperature predictions can be generalized inside these regions.

In addition to the target domains of this study, the predictabilities of the Mediterranean region and eastern Europe were tested. They were found to be lower than in SC and WE for most of the seasons, sometimes considerably weaker. Results from those experiments were not included in this paper, because many conclusions are based on averaging the results over all domains, and the risk for misleading conclusions would be increased if less predictable domains were included. On the other hand, this means that our results primarily apply to those parts of Europe where the prediction skill is highest.

d. Random sampling of the training data

After preprocessing of the predictor and predictand data and derivation of the potential predictors using all years of the study, the years 1986–2010 were put aside to be used later as an independent test sample. These 25 years did not participate in fitting of the models, in the predictor selection procedure, or in the calibration of the post-processing model.

The training data, consisting of years 1915–1985 and 31 potential predictors, were then sampled 1000 times so that in each iteration, 35 years (50% of all fitting years) and 10 predictors (33% of all predictors, the percentage value suggested by Hastie et al. (2009) for regression problems), were selected randomly without replacement to be used in fitting of each individual LASSO regression model. Building a model ensemble by this way is called random subspaces sampling, or attribute bagging (e.g., Bryll et al. 2003) and it is closely related to the bootstrap-aggregation (i.e., bagging; Hastie et al. 2009). In the attribute bagging approach the data samples are smaller than in bagging, which speeds up the calculation considerably, and also increases the accuracy of the ensemble by decreasing the correlation among the individual regression models.

e. LASSO regression in random samples

For each (1) target domain, (2) season, and for each (3) random sample, one LASSO regression model was fitted using Python's Scikit-learn library (version 0.20.1; Pedregosa et al. 2012). Automatic predictor selection and shrinkage of the sampled predictors was applied using cross-validation and the least angle regression algorithm (LARS; Efron et al. 2004) as implemented in the LassoLarsCV function.

LASSO differs from the ordinary least squares regression (OLS) such that the quantity to be minimized,

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=i}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}\left|\beta_j\right| \qquad = \text{RSS} + \mathit{penalty}\,, \tag{1}$$

where i denotes the sample and j the predictor indices, y the predictand, and β the regression coefficient, contains an extra penalty term, $\lambda \sum_{j=1}^{p}\left|\beta_j\right|$, in addition to the residual sum of squares, RSS = $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=i}^{p}\beta_j x_{ij})^2$, minimized in OLS (Hastie et al. 2009). The extra term defines the amount of shrinkage of the regression coefficients β_j, controlled by the regularization parameter $\lambda \geq 0$. Depending on the value of λ, the coefficients β_j may be positive, negative, or zero, the latter leading effectively to rejection of predictors.

In this study, K-fold cross-validation with K = 5 was used to optimize the value of λ for each ensemble member. The algorithm first fits a separate LASSO model for each K fold, and then calculates a separate regularization path and the related mean square error for all of them. Finally the optimal value for the parameter is selected by first calculating the mean of mean square errors of all K solutions, and then investigating the minimum of it. Using K = 20 was also tested, but it was found to slow down the fitting process without improving the result.

The LARS algorithm is essentially the same as forward stagewise predictor selection, which typically yields a higher number of selected predictors than other, greedier approaches, such as forward selection. Despite that, the weakest predictors were either rejected or considerably regularized in the fitting, as shown in section 3c. This regularization of the β_j coefficients decreases the risk of overfitting of the individual ensemble members compared to using, for example, the unregularized OLS.

f. Deterministic validation metrics

Two metrics were selected for validation of the quality of modeling against observed values, the most important being the anomaly correlation coefficient (ACC; Wilks 2011, section 8.6), referring to the Pearson correlation coefficient calculated between the observed and modeled predictand anomaly time series. Additionally, the mean square skill score was used, defined as

$$MSSS_{ref} = 1 - \frac{\frac{1}{n}\sum_{t=1}^{n}(y_m - y_o)^2}{\frac{1}{n}\sum_{t=1}^{n}(y_{ref} - y_o)^2} = 1 - \frac{MSE_m}{MSE_{ref}} \quad . \qquad (3)$$

The y_m and y_o symbols denote the modeled and observed time series respectively. As reference forecasts, denoted by y_{ref}, the climatological and the persistence forecasts were used, denoted $MSSS_{clim}$ and $MSSS_{pers}$, assuming zero anomalies always for the former, and maintained observed anomalies from one season to the next for the latter.

The ACC value describes the model ability to reproduce the anomalies (independently of their amplitude), while $MSSS$ tells about the model's relative accuracy compared to the given reference forecast. ACC and $MSSS$ can have values in the range [-1, 1] and [-∞, 1], respectively. Both scores are positively oriented, so that the higher the score, the better the result. In general, negative or near-zero ACC values indicate non-skillful forecasts, when estimated from an independent test sample. Positive $MSSS$ values indicate that the forecast is able to surpass the skill of the reference.

All metrics were calculated using the ensemble mean of the LASSO models as a deterministic forecast, because those metrics are fast to calculate and interpret, and are suitable for model developing purposes. Probabilistic interpretation of the ensemble results remains to be applied in further work. To some extent the deterministic performance of the ensemble mean also reflects the probabilistic performance: it is unlikely that an inaccurate deterministic model would produce valuable probabilistic predictions, but an accurate one might well do so.

g. Post-processing the raw output

After creating the ensemble of the LASSO models, each member of the ensemble was applied to predict the whole 1915–2010 period, including also the test years. Minimizing the RSS in the fitting of the LASSO regression models typically leads to reduced variance in the model output, when compared to the observed variability of the predictands. This was handled by increasing the variability of the modeled predictand values by a non-linear quantile mapping method (modified from the approach of Räisänen and Räty, 2013), such that results from each ensemble member were corrected separately using a model-specific quantile correction function

$$y_c = F_o^{-1}(F_m(y_m)) \quad , \qquad (2)$$

where y_c represents the corrected predictand time series, y_m the raw model output, F_m the cumulative distribution function calculated from the raw model data, and F_o^{-1} the corresponding inverse distribution from observed values. The correction functions were calculated for all models separately using the training years, and then applied to correct all years 1915–2010. Additionally, the same correction was applied to the persistence reference forecasts, because typically the variability differs quite a lot between seasons, which would make $MSSS_{pers}$ too high if not using the correction.

3. Results

a. Prediction skill of temperatures in Europe

Temperature predictability in all domains and seasons was estimated using the validation metrics, calculated over the independent test years. As seen in Figure 5, the ACC of the ERA-20C based ensemble is positive and statistically significant in all domains and seasons except in DJF and MAM in WE. The mean ACC, calculated over all seasons and domains, is 0.51. Essentially the same mean ACC score, 0.49, was achieved when the 20CRv2c reanalysis was used to train and test the model ensemble, indicating that both reanalyses contain information that can be extracted by the PCA and used in fitting and forecasting. On average, also the skill scores of the individual domains and seasons are comparable between the two reanalyses. In general, the prediction skill of temperatures in the JJA and SON seasons is better than in DJF and MAM, and prediction skill of the largest EU domain is higher than that in smaller domains.

The performance in terms of the $MSSS$ scores is not as good as that measured by the ACC when comparing the number of statistically significant target domains and seasons. The $MSSS$ scores can be decomposed to different terms, where ACC is one term (Murphy and Epstein 1989), and ACC is therefore a measure of *potential* rather than *actual* skill. Overcoming the persistence forecast, for example, can be difficult because of the often strong autocorrelation of the predictand. Autocorrelation brings ACC skill to the persistence reference forecast, as shown in the bottom-right legends of Figure 6, and for this reason it is difficult to overcome by the LASSO ensemble. Despite that, the climatological forecast seems to be even more difficult to beat, seen in the lower mean $MSSS_{clim}$ value (0.24) compared to the mean $MSSS_{pers}$ (0.37).

b. Sensitivity of skill to details of methodology

Including the local persistence of the predictand variable into the list of the potential predictors was found to improve the accuracy of the modeling such that the mean ACC would increase from 0.51 to 0.54, enhancing the prediction skill especially in the WE region. Using the persistence of the predictand as a predictor is common in different studies (e.g., Karpechko 2015; Kolstad et al. 2015; Johansson et al. 1998), but as the main objective of this work was to study and show the potential of lagged SST and GPT PCs, it was not further used here.

According to modeling tests with different domains and reanalyses, temperature forecasts are the most skillful and consistent when the global domain is used for extraction of SST PCs, compared to using the smaller Northern Hemispheric domain. The predictive power of GPT, however, was stronger when signals from the Northern Hemisphere only were taken into account. To test the relative importance of the SST and GPT input parameters, both were excluded in the modeling in turn. When GPT was excluded, the mean ACC decreased from 0.51 to 0.44 (from 0.49 to 0.44 in 20CRv2c) and when SST was excluded, the mean ACC decreased again to 0.44 (0.39 in 20CRv2c). In general, SST performs better in JJA and SON, and GPT in DJF and MAM. These results show that both input parameters could be used in seasonal forecasting separately, and both bring equally much predictive power in the forecasts on average. The best result, however, was achieved when both parameters were included.

Using the quantile mapping correction for individual ensemble members is beneficial for the $MSSS$ skill in those domains and seasons where the prediction skill is high enough prior to correction (ACC > 0.40), but it does not improve results significantly, if at all, if the original, uncorrected results are not very skillful. Without correction the variance of the ensemble mean would be near zero, which explains the improvement in $MSSS$ after the correction. Because of the near-zero variance of the ensemble mean, the simulated trends were also improved considerably. The effect on the ACC score was only minor, even though the non-linear correction could, in principle, affect also that metric.

The usefulness of predictor lagging was tested by excluding lags 2–5 from the potential predictors, such that only predictor values from the previous season were used to forecast the next season. As a consequence, the mean ACC dropped from 0.51 to 0.46, indicating that using lagged predictors is valuable. Similarly, the importance of using long time series in fitting and predictor selection was tested by excluding the period 1915–1944 from the fitting and predictor selection years. The mean ACC dropped even more, to 0.41, probably because of increase in the ratio $\frac{p}{n}$ which leads to overfitting, or because of increased sampling uncertainty related to the preprocessing: calculation of trends, climatology, and PCs require a certain amount of data to be robust.

A large part of the predictability seems to arise from the correct simulation of the decadal variability and decadal trends. For example, in SC the observed and modeled temperature trends are positive during the test years in all seasons, but especially in DJF and SON, as seen in Figure 6. To study the effect of decadal trend on the ACC, linear trends were removed separately from the modeled and observed test period time series. It was found that the mean ACC decreased from 0.51 to 0.36.

The accuracy of the ensemble mean was compared to the accuracy of the best individual ensemble member for all domains and seasons. To find the best member, the ACC of each LASSO model was calculated for the fitting period 1915–1985, and the member with the highest score was then selected. It was found that the mean ACC in test years was only 0.34 for the best individual members, showing the power of the ensemble mean approach. As Figure 7b shows, including more than 50 members in the ensemble does not improve the ensemble mean accuracy very often. However, growing the ensemble size larger than that is not usually harmful either, and sometimes the skill peaks with considerably larger ensemble sizes, which justifies the selection of using 1000 members in this study.

The evolution of ACC over all years 1915–2010 was also analyzed by applying a moving 25-year analysis window. As shown in Figure 7a, the temporal variations in ACC were typically small and the absolute minimum values high in the JJA and SON seasons over all years, in addition to their better test period skill compared to that of DJF and MAM. Somewhat larger but slowly changing variability in the ACC skill was discovered between different periods in DJF and MAM seasons. Weisheimer et al. (2017) reported a corresponding time-dependency of skill in their winter NAO forecasts, and found, for example, that their forecast skill drops around 1960 and increases again later. In our analysis the timing of high and low skill periods depends on the season and domain, but on average, the skill seems to be lower around 1950–1970. The skill does not drop when crossing the boundary between the fitting and testing years, but remains essentially the same. This shows that the model ensemble is not overfitted, and thus builds confidence to the true skill of the statistical modeling.

c. Analysis and interpretation of the predictors in the ensemble members

The predictors in the LASSO models differ depending on the season and domain. Additionally the number of predictors in the models varies, as can be seen in the analysis of predictors in Figure 8. Typically a large average number of predictors was found in SON and sometimes in the JJA seasons, and models fitted to forecast DJF and MAM contained fewer predictors.

Different PCs of GPT were found in all seasons, but they were especially common in the MAM models. While the second PC of GPT, GPT2, is by far the most common in all domains and seasons, all lags of SST2 in no particular order are in the top 15 among the 31 potential predictors. Temporally remote signals from past seasons appear in the models sometimes, as lags of three to five seasons were found for different PCs of SST. Lags 4 and 5, however, tend not to be as common as lags 1–3. Interestingly, the dominant components of both parameters, SST1 and GPT1, were both quite rare among the ensemble members. Forcing the ENSO-related SST1 component manually to be included in all random samples did not improve the results. In other studies the ENSO has been found to affect, for example, the NAO variability (e.g., Hall et al. 2017; Bell et al. 2009) and the European climate in general

(Brönnimann 2007). It is likely that the response to the ENSO forcing is non-linear in Europe, and to be able to utilize SST1, linear regression models would require major modifications applied to it, as shown by Folland et al. (2012) and Hall et al. (2017). Preliminary tests of using the ENSO signal modification of Folland et al. did not bring any useful predictive power to the modeling, though, probably because of using seasonal mean values as predictors instead of monthly means.

On the other hand, the SST2 component, mostly describing the AMO on decadal scales, was the most frequent SST predictor on average. This result not only indicates the strong role of the North Atlantic in regulating the climate of Europe (Li et al. 2018; Chen et al. 2018; Knight et al. 2006; Czaja and Frankignoul 2002; Peng et al. 2005), but also reveals the importance of decadal scale SST modes as a source of predictability in seasonal forecasts (Davini et al. 2015). Out of all SST components, only SST2 contains a clear trend during the test period years, which overlaps with the temperature trends in different predictand seasons. This trend partly explains the correct simulation of the decadal temperature trends during the test period years, especially in JJA and SON seasons. Similar decadal scale variability and trends can be seen in the GPT predictors, and they contribute also to the simulation of decadal trends in the predictands.

Finally, SST3 and SST4, which primarily describe different shorter-term variability modes of the Pacific Ocean (as well as the PDO), are relatively common in all seasons, but especially in DJF and MAM. Corresponding shorter scale variability is also visible in SST2, in addition to the quite strong decadal scale signal of it. All lags of SST2 were common among the models, which might indicate that the predictive power of that component originates from the decadal, low frequency AMO variability, and not from the higher frequencies of it (i.e., annual or seasonal scale variability). However, whether the longer-term variability of SST2 actually is more important than its shorter-scale variability in bringing prediction skill to seasonal forecasts, or vice versa, remains to be studied in further work.

4. Discussion

Predictor selection is a sensitive and potentially error-prone task, which can fail if the quality of the predictand or predictors is low, if the internal variability of the predictand is prominent compared to the predictable component, and if the number of samples is too small. The selected modeling approach could be used directly to forecast grid cell values, but that would be risky, because the internal variability of them is probably too large. Too large internal variability may cause problems in predictor selection and regularization. In this study the predictands derived from the smaller target domains might still contain too much internal variability despite the areal averaging, which could explain at least partly why the quality of the results in WE was occasionally lower than in the other domains. A possible cure for this problem could be to use those predictors that were found from the largest EU domain to predict the smaller domains, using refitted models. In the EU domain the signal-to-noise ratio is higher and predictor selection might work better there for that reason, and it is possible that the same global predictors that bring predictability to the largest domain also work for the smaller domains. On the other hand, the risk of this approach is that if the target domain is too large, different parts of it should be predicted using partly or completely different predictors.

However, this avenue of exploration is left for further research.

Another potential risk in the predictor selection is related to using lagged, occasionally highly collinear predictors. Collinearity refers to the correlations between predictors, and it is typically high between differently lagged versions of the same predictor. Autocorrelation analysis of the SST PCs, for example, indicates that the five lags are positively correlated in all five components (not shown). Collinearity can increase the probability of selecting non-optimal lags. Even though the skill of the models might not degenerate very much from the use of non-optimal lags, compared to using too many and/or too weak potential predictors, the explicability and interpretability of the models can weaken.

However, according to the modeling results, the potential drawbacks of using lagged predictors seem to be minor compared to the benefits of it. Because of the very complex and possibly chained interactions between different components of the climate system (Vihma 2014), it is possible that some previously known or unknown mechanisms increase the prediction skill in the modeling. Propagation of the signal through such known pathways as SST→GPT→troposphere (Bell et al. 2009; Kidston et al. 2015), and continental snow cover→GPT→troposphere (Gastineau et al. 2017) can take several months or seasons, which probably explains why using different lags was beneficial. At best, statistical learning can find and take the advantage of these mechanisms to improve the accuracy of the forecasts.

Each random sample contained only 50% of the fitting years. Those years were used to fit one LASSO model, and the remaining 50% could be used in validation of that particular model. Further, the validation years could be used to weight the models based on their validation ACC skill in order to enhance the accuracy of modeling. This procedure was briefly tested, but it was found that the accuracy of the ensemble mean was not better compared to using unweighted models. It is possible that fittings performed on some certain subset of fitting years produce models that perform well in the validation sample, and when the number of random samples is high enough, models that were fitted using roughly that particular set of years get heavier weights. Effectively, this leads to rejection of some potentially valuable fitting years. For successful weighting of models, probably more data (i.e., longer time series), would be needed.

The earliest years of the century are more uncertain in the SST predictor data, due to sparseness of the observations at that time (Kennedy 2013), and the same time-dependency of uncertainty also holds for the GPT data. These uncertainties affect the fitting of the LASSO models. However, if the biases of the predictor data are not systematic over the years, but random and not too large, they might not hamper the fitting of the models too much. This hypothesis is supported by the test where exclusion of the earliest fitting years was found to decrease the accuracy of modeling (section 3b). The uncertainty of the most recent years, which were selected for testing purposes, is anyhow smaller, probably increasing the validation score values in those years to some extent. On the other hand, the temperatures in Europe are probably comparatively well described in the HadCRUT4 data throughout the years, thanks to the dense weather station network there throughout the century, reducing the uncertainty related to the predictands.

A large part of existing literature of seasonal predictability concentrates on the wintertime NAO index forecasts (section 1), and because it was not a predictand of this study, comparing the skill of our modeling to that of others is challenging. References to summer or fall temperature forecasting experiments are especially sparse. Further, the measures of skill vary between different studies, and comparing, for example, the probabilistic or categorical skill scores used in many papers (Karpechko, 2015; Weisheimer and Palmer 2014; Graham et al. 2005) with our deterministic scores is difficult. However, regarding the mostly insignificant, and occasionally even negative ACC skill of temperature forecasts of the current operational forecasting models (Mishra et al. 2018), it is probably safe to say that dynamical models perform worse than our model at least in summer and fall. In the statistical, pre-reanalysis approach of Colman and Davey (1999), a good but not completely independently estimated summer predictability was found for a region which roughly corresponds to our WE domain. Interestingly, they used North Atlantic SSTs as predictors, which were also an important source of predictability in our modeling. Johansson et al. (1998), as well, studied the predictability of seasonal mean temperatures in northern Europe, using PCAs of SST as predictors, and concluded that the fall season was the least predictable in their approach. In our modeling that season was the most predictable. They found most skill in winter forecasts, but even for that season their ACC score was not very high.

In our approach some potentially important predictors, such as sea ice and snow cover could not be used because they were too uncertain over the period of our study (Appendix A).

Accurate enough predictors derived from these parameters might have improved the slightly poorer winter and spring forecasts, as the variability and predictive power of these parameters are strongly connected to those seasons (e.g., Liptak and Strong 2014). Preliminary tests suggest also that using monthly mean values for predictor parameters, instead of seasonal means, improves the accuracy of the forecasts in some seasons and domains, especially in spring. That season is especially sensitive to the GPT variability, which partly explains the improvement: the wintertime SSW events typically affect the spring the most, and apparently, identifying them correctly from seasonally averaged data can be too tricky for the regression models. Further, seasonally averaged predictor data give no information about the timing of the events within the three-month winter season, which is important because of the relatively short time frame of the stratosphere-troposphere interaction (e.g., Baldwin and Dunkerton 2001).

Replacing the ERA-20C and the 20CRv2c datasets with some other reanalyses would have improved the reliability of many or all input parameters, probably to the extent where, for example, predictors derived from the sea ice would have been applicable. However, other reanalyses, such as ERA-Interim, usually contain considerably shorter time series. Increasing the number of potential predictors p and at the same time decreasing the number of fitting years n would both increase the dimensionality of the data (i.e., the ratio $\frac{p}{n}$) , which most likely would lead to overfitting. In other words, using long time series for training the model ensemble allows us to use more PCs and more lags per predictor parameter, such that the full potential of the parameters can be utilized, probably better than would be possible with shorter time series.

5. Conclusions

In this paper, we present a new method to forecast seasonal temperatures in Europe. The skill of the model was estimated, and the most important predictors were identified to explain the sources of predictability. The main findings are collected below.

- Statistical learning is useful in the field of seasonal forecasting. Its complexity is between the very simple statistical OLS models and heavy dynamical models. The selected ensemble approach of this study is suitable for smaller datasets, and is reasonably tractable in physical interpretation compared to other, more flexible and highly nonlinear approaches, such as neural networks or support vector machines (Hastie et al. 2009). The statistical learning properties of the method can both help in selecting and weighting the most important predictors, and in improving the accuracy of the results, when compared to the skill of the best individual ensemble member.
- As already shown in previous papers, Northern Hemispheric GPT variability in preceding seasons was confirmed to bring predictability into forecasts of European temperatures over various domains, mostly in winter and spring.
- When compared to GPT, the global SST variability, including the PDO and most importantly the AMO indices, is an equally important source of predictability on average. Because of long enough time series used for fitting of models, decadal variability was captured and then skillfully exploited during the test period years. Summer and fall forecasts benefit the most of using the SST as a predictor.
- Using lagged SST predictors increases the predictive power of the statistical model

ensemble compared to the more traditional approach, where only values from the previous season, or month, are used. Slowly evolving currents and other processes in the oceans contain information and signals that can be exploited by using predictor values from further past.

- Two centennial scale reanalyses, ERA-20C and 20CRv2c, were used separately in derivation of the potential predictors. The prediction skill was found to be insensitive to the choice of the reanalysis product. Additionally, the predictive skill was roughly on the same level in the fitting and testing periods on average. These results indicate the robustness and low variance of the model system.
- Compared to the skill achieved in previous papers on seasonal temperature forecasts for European target regions, the forecasts for summer and fall were especially skillful and robust.

Acknowledgements

MK was supported by the Kone Foundation, via a personal PhD grant for studying the climatic effects of ocean surface temperature in the Northern Hemisphere (https://koneensaatio.fi/en/grants/tuetut/2016-2/; grant 088575), which was the main source of funding of this study. PU was supported by the EC Marie Curie Support Action LAWINE (grant 707262). AYK was supported by the Academy of Finland (grants 286298, 319397). OH and IL were supported by the Finnish Meteorological Institute. JR was supported by the Academy of Finland Centre of Excellence program (project 307331).

The Python code for reproducing the results of this study is available online at https://github.com/fmidev/seasonal_forecasting.

Appendix

APPENDIX A

Exclusion of sea ice and snow cover variables from predictor parameters

In previous studies the potential of the sea ice cover or concentration (SIC) as a source of predictability was recognized (e.g., Bader et al. 2011). Unfortunately, appropriate enough sea ice predictor data could not be found for this study, as the longest available sea ice records contain many uncertainties, which additionally vary in time. Attempts to decompose the HadISST1 (Rayner et al. 2003) or Walsh (2015) sea ice data sets with PCA were not successful: the discontinuities and changes in the observational systems could be clearly seen in the derived PCs, making fitting and application of the LASSO models too uncertain. According to Bunzel et al. (2016), dynamical seasonal models are sensitive to details in the sea ice, which not only highlights the importance of ice concentration data for seasonal forecasting, but also reveals the risks related to using too uncertain datasets for fitting of statistical models.

The PCs of snow cover (SNC) were found to be more homogeneous in reanalyses than PCs of SIC in observational datasets. Using PCs of SNC as predictors was found to be possible and beneficial for the accuracy in some cases, and thus they could be used as predictors in seasonal forecasting. However, due to following reasons, they were excluded from this study:

- When SNC was used, the results were slightly less accurate than with other predictors. Combining SNC with SST and GPT predictors did not improve the model skill compared to using SST and GPT only.
- The snow covers in ERA-20C and 20CRv2c were found to differ, and less skillful results were achieved when 20CRv2c SNC was used as a predictor parameter, compared to using ERA-20C SNC.
- Previous studies have revealed inaccuracies in the reanalysis snow data (e.g., Wegmann et al. 2016).

APPENDIX B

Connection of SST principal components to known SST indices

According to Messié and Chavez (2011), the leading five or six SST PCs can be associated with the major variability modes of the world ocean, namely the El Niño–Southern Oscillation (ENSO), the Atlantic Multidecadal Oscillation (AMO), the Pacific Decadal Oscillation (PDO), the North Pacific Gyre Oscillation (NPGO), El Niño Modoki index (EMI), and the Atlantic El Niño index (ATL3). The results of Messié and Chavez, namely the monthly principal components of SST, were downloaded from http://climexp.knmi.nl/, season-averaged, and then compared to our components using cross-correlation analysis. Strong correlations were found between the first components (representing ENSO; $r = 0.98$), and the second components (AMO; $r = 0.80$). The third and fourth components (representing mostly the Pacific modes PDO, NPGO, and EMI) were again strongly correlated ($r = -0.79$ and

$r = -0.72$, respectively). The fifth (EMI, ATL3) component is not very strongly correlated between the two PC datasets ($r = 0.33$), and it leaks some variability also to the sixth component ($r = 0.56$). Differences between the datasets arise from the use of different SST data, different time periods and resolution, and partly different preprocessing steps.

References

*Allen, R. J., and C. S. Zender, 2011: Forcing of the Arctic Oscillation by Eurasian snow cover. J. Clim., **24**, 6528–6539, doi:10.1175/2011JCLI4157.1.*

*Bader, J., M. D. S. Mesquita, K. I. Hodges, N. Keenlyside, S. Østerhus, and M. Miles, 2011: A review on Northern Hemisphere sea-ice, storminess and the North Atlantic Oscillation: Observations and projected changes. Atmos. Res., **101**, 809–834, doi:10.1016/j.atmosres.2011.04.007. http://dx.doi.org/10.1016/j.atmosres.2011.04.007.*

*Baker, L. H., L. C. Shaffrey, and A. A. Scaife, 2018: Improved seasonal prediction of UK regional precipitation using atmospheric circulation. Int. J. Climatol., **38**, e437–e453, doi:10.1002/joc.5382.*

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., 7808–7817, doi:10.1029/2018GL078838. http://doi.wiley.com/10.1029/2018GL078838.

*Baldwin, M. P., and T. J. Dunkerton, 2001: Stratospheric harbingers of anomalous weather regimes. Science (80-.)., **294**, 581–584, doi:10.1126/science.1063315.*

*Bell, C. J., L. J. Gray, A. J. Charlton-Perez, M. M. Joshi, and A. A. Scaife, 2009: Stratospheric communication of El Niño teleconnections to European winter. J. Clim., **22**, 4083–4096, doi:10.1175/2009JCLI2717.1.*

*Brönnimann, S., 2007: Impact of El Niño–Southern Oscillation on European climate. Rev. Geophys., **45**, doi:doi:10.1029/2006RG000199.*

*Brönnimann, S., and Coauthors, 2016: Multidecadal variations of the effects of the Quasi-Biennial Oscillation on the climate system. Atmos. Chem. Phys., **16**, 15529–15543, doi:10.5194/acp-16-15529-2016.*

*Bryll, R., R. Gutierrez-Osuna, and F. Quek, 2003: Attribute bagging: Improving accuracy of classifier ensembles by using random feature subsets. Pattern Recognit., **36**, 1291–1302, doi:10.1016/S0031-3203(02)00121-8.*

*Bunzel, F., D. Notz, J. Baehr, W. A. Müller, and K. Fröhlich, 2016: Seasonal climate forecasts significantly affected by observational uncertainty of Arctic sea ice concentration. Geophys. Res. Lett., **43**, 852–859, doi:10.1002/2015GL066928.*

Chen, S., R. Wu, W. Chen, and S. Yao, 2018: Enhanced linkage between Eurasian winter and spring dominant modes of atmospheric interannual variability since the early-1990s. J. Clim., JCLI-D-17-0525.1, doi:10.1175/JCLI-D-17-0525.1. http://journals.ametsoc.org/doi/10.1175/JCLI-D-17-0525.1.

*Clark, R. T., P. E. Bett, H. E. Thornton, and A. A. Scaife, 2017: Skilful seasonal predictions for the European energy industry. Environ. Res. Lett., **12**, doi:10.1088/1748-9326/aa57ab.*

*Cohen, J., and D. Entekhabi, 1999: Eurasian snow cover variability and Northern Hemishpere climate predictability. Geophys. Res. Lett., **26**, 345–348,* *doi:10.1029/1999GL900200.*

Cohen, J., and J. Jones, 2011: A new index for more accurate winter predictions. Geophys. Res. Lett., ***38****, 1–6, doi:10.1029/2011GL049626.*

Colman, A., and M. Davey, 1999: Prediction of summer temperature, rainfall and pressure in Europe from preceding winter North Atlantic Ocean Temperature. Int. J. Climatol., ***19****, 513–536, doi:10.1002/(SICI)1097-0088(199904)19:5<513::AID-JOC370>3.0.CO;2-D.*

Compo, G. P., and Coauthors, 2011: The Twentieth Century Reanalysis Project. Q. J. R. Meteorol. Soc., ***137****, 1–28, doi:10.1002/qj.776.*

Czaja, A., and C. Frankignoul, 2002: Observed impact of Atlantic SST anomalies on the North Atlantic oscillation. J. Clim., ***15****, 606–623, doi:10.1175/1520-0442(2002)015<0606:OIOASA>2.0.CO;2.*

Davini, P., J. Von Hardenberg, and S. Corti, 2015: Tropical origin for the impacts of the Atlantic Multidecadal Variability on the Euro-Atlantic climate. Environ. Res. Lett., ***10****, doi:10.1088/1748-9326/10/9/094010.*

De Cian, E., E. Lanzi, and R. Roson, 2013: Seasonal temperature variations and energy demand. Clim. Change, ***116****, 805–825, doi:10.1007/s10584-012-0514-5. http://link.springer.com/10.1007/s10584-012-0514-5.*

Dee, D. P., and Coauthors, 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.*

DelSole, T., and J. Shukla, 2009: Artificial skill due to predictor screening. J. Clim., ***22****, 331–345, doi:10.1175/2008JCLI2414.1.*

Dobrynin, M., and Coauthors, 2018: Improved Teleconnection-Based Dynamical Seasonal Predictions of Boreal Winter. Geophys. Res. Lett., doi:10.1002/2018GL077209.

Dunstone, N., and Coauthors, 2018: Skilful Seasonal Predictions of Summer European Rainfall. Geophys. Res. Lett., doi:10.1002/2017GL076337.

*Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani, 2004: Least Angle Regression. Ann. Stat., **32**, 407–499.*

*Folland, C. K., A. A. Scaife, J. Lindesay, and D. B. Stephenson, 2012: How potentially predictable is northern European winter climate a season ahead? Int. J. Climatol., **32**, 801–818, doi:10.1002/joc.2314.*

*Förster, K., and Coauthors, 2018: Retrospective forecasts of the upcoming winter season snow accumulation in the Inn headwaters (European Alps). Hydrol. Earth Syst. Sci., **22**, 1157–1173, doi:10.5194/hess-22-1157-2018.*

*Gastineau, G., J. García-Serrano, and C. Frankignoul, 2017: The influence of autumnal Eurasian snow cover on climate and its link with Arctic sea ice cover. J. Clim., **30**, 7599–7619, doi:10.1175/JCLI-D-16-0623.1.*

*Gerber, E. P., and P. Martineau, 2018: Quantifying the variability of the annular modes: Reanalysis uncertainty vs. sampling uncertainty. Atmos. Chem. Phys., **18**, 17099–17117, doi:https://doi.org/10.5194/acp-18-17099-2018. www.atmos-chem-phys.net/18/17099/2018/.*

Graham, R. J., M. Gordon, P. J. McLean, S. Ineson, M. R. Huddleston, M. K. Davey, A. Brookshaw, and R. T. H. Barnes, 2005: A performance comparison of coupled and uncoupled versions of the Met Office seasonal prediction general circulation model.

Tellus, Ser. A Dyn. Meteorol. Oceanogr., ***57****, 320–339, doi:10.1111/j.1600-0870.2005.00116.x.*

Hall, R. J., A. A. Scaife, E. Hanna, J. M. Jones, and R. Erdélyi, 2017: Simple Statistical Probabilistic Forecasts of the Winter NAO. Weather Forecast., ***32****, 1585–1601, doi:10.1175/WAF-D-16-0124.1. http://journals.ametsoc.org/doi/10.1175/WAF-D-16-0124.1.*

Hastie, T., R. Tibshirani, and J. Friedman, 2009: The Elements of Statistical Learning. Second Edi. Springer, New York, 1-745 pp. http://www.web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf.

Huang, J., H. M. Van den Dool, and A. G. Barnston, 1996: Long-lead seasonal temperature prediction using optimal climate normals. J. Clim., ***9****, 809–817.*

Jia, L., and Coauthors, 2017: Seasonal prediction skill of northern extratropical surface temperature driven by the stratosphere. J. Clim., ***30****, 4463–4475, doi:10.1175/JCLI-D-16-0475.1.*

Johansson, Å., A. Barnston, S. Saha, and H. van den Dool, 1998: On the Level and Origin of Seasonal Forecast Skill in Northern Europe. J. Atmos. Sci., ***55****, 103–127, doi:10.1175/1520-0469(1998)055<0103:OTLAOO>2.0.CO;2.*

Karpechko, A. Y., 2015: Improvements in statistical forecasts of monthly and two-monthly surface air temperatures using a stratospheric predictor. Q. J. R. Meteorol. Soc., ***141****, 2444–2456, doi:10.1002/qj.2535. http://doi.wiley.com/10.1002/qj.2535.*

Karpechko, A., K. Andrew Peterson, A. A. Scaife, J. Vainio, and H. Gregow, 2015: Skilful seasonal predictions of Baltic Sea ice cover. Environ. Res. Lett., ***10****, doi:10.1088/1748-9326/10/4/044007.*

Kennedy, J. J., 2013: A review of uncertainty in in situ measurements and data sets of sea surface temperature. Rev. Geophys., ***52****, 1–32, doi:10.1002/2013RG000434.*

Kidston, J., A. A. Scaife, S. C. Hardiman, D. M. Mitchell, N. Butchart, M. P. Baldwin, and L. J. Gray, 2015: Stratospheric influence on tropospheric jet streams, storm tracks and surface weather. Nat. Geosci., ***8****, 433–440, doi:10.1038/NGEO2424. http://dx.doi.org/10.1038/ngeo2424.*

Knight, J. R., C. K. Folland, and A. A. Scaife, 2006: Climate impacts of the Atlantic multidecadal oscillation. Geophys. Res. Lett., ***33****, doi:10.1029/2006GL026242.*

Kolstad, E. W., and M. Årthun, 2018: Seasonal Prediction from Arctic Sea Surface Temperatures: Opportunities and Pitfalls. J. Clim., JCLI-D-18-0016.1, doi:10.1175/JCLI-D-18-0016.1. http://journals.ametsoc.org/doi/10.1175/JCLI-D-18-0016.1.

Kolstad, E. W., S. P. Sobolowski, and A. A. Scaife, 2015: Intraseasonal persistence of European surface temperatures. J. Clim., ***28****, 5365–5374, doi:10.1175/JCLI-D-15-0053.1.*

Li, F., Y. J. Orsolini, H. Wang, Y. Gao, and S. He, 2018: Atlantic Multidecadal Oscillation Modulates the Impacts of Arctic Sea Ice Decline. Geophys. Res. Lett., ***45****, 2497–2506, doi:10.1002/2017GL076210.*

Liptak, J., and C. Strong, 2014: The winter atmospheric response to sea ice anomalies in the barents sea. J. Clim., ***27****, 914–924, doi:10.1175/JCLI-D-13-00186.1.*

Messié, M., and F. Chavez, 2011: Global modes of sea surface temperature variability in relation to regional climate indices. J. Clim., ***24****, 4314–4331, doi:10.1175/2011JCLI3941.1.*

Mishra, N., C. Prodhomme, and V. Guemas, 2018: Multi-Model Skill Assessment of Seasonal Temperature and Precipitation Forecasts over Europe. Clim. Dyn., ***0****, 29–31, doi:10.1007/s00382-018-4404-z. http://dx.doi.org/10.1007/s00382-018-4404-z.*

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. J. Geophys. Res., ***117****, doi:doi:10.1029/2011JD017187.*

Murphy, A. H., and E. S. Epstein, 1989: Skill Scores and Correlation Coefficients in Model Verification. Mon. Weather Rev., ***117****, 572–581.*

Orth, R., and S. I. Seneviratne, 2014: Using soil moisture forecasts for sub-seasonal summer temperature predictions in Europe. Clim. Dyn., ***43****, 3403–3418, doi:10.1007/s00382-014-2112-x.*

Ossó, A., R. Sutton, L. Shaffrey, and B. Dong, 2017: Observational evidence of European summer weather patterns predictable from spring. Proc. Natl. Acad. Sci., 201713146, doi:10.1073/pnas.1713146114. http://www.pnas.org/lookup/doi/10.1073/pnas.1713146114.

Pedregosa, F., and Coauthors, 2012: Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res., ***12****, 2825–2830, doi:10.1007/s13398-014-0173-7.2. http://dl.acm.org/citation.cfm?id=2078195\nhttp://arxiv.org/abs/1201.0490.*

Peng, S., W. A. Robinson, S. Li, and M. P. Hoerling, 2005: Tropical Atlantic SST forcing of coupled North Atlantic seasonal responses. J. Clim., ***18****, 480–496, doi:10.1175/JCLI-3270.1.*

*Poli, P., and Coauthors, 2016: ERA-20C: An atmospheric reanalysis of the twentieth century. J. Clim., **29**, 4083–4097, doi:10.1175/JCLI-D-15-0556.1.*

*Räisänen, J., and O. Räty, 2013: Projections of daily mean temperature variability in the future: cross-validation tests with ENSEMBLES regional climate simulations. Clim. Dyn., **41**, 1553–1568, doi:10.1007/s00382-012-1515-9. http://link.springer.com/10.1007/s00382-012-1515-9.*

*Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J. Geophys. Res., **108**, doi:10.1029/2002JD002670. http://doi.wiley.com/10.1029/2002JD002670.*

*Rodwell, M. J., and C. K. Folland, 2002: Atlantic air – sea interaction and seasonal predictability. Q. J. R. Meteorol. Soc., **128**, 1413–1443, doi:https://doi.org/10.1002/qj.200212858302.*

*Scaife, A. A., and Coauthors, 2016: Seasonal winter forecasts and the stratosphere. Atmos. Sci. Lett., **17**, 51–56, doi:10.1002/asl.598.*

*Scaife, A., and Coauthors, 2014: Skillful long range prediction of European and North American winters. Geophys. Res. Lett., **5**, 2514–2519, doi:10.1002/2014GL059637. http://onlinelibrary.wiley.com/doi/10.1002/2014GL059637/full.*

*Smith, D. M., A. A. Scaife, R. Eade, and J. R. Knight, 2016: Seasonal to decadal prediction of the winter North Atlantic Oscillation: Emerging capability and future prospects. Q. J. R. Meteorol. Soc., **142**, 611–617, doi:10.1002/qj.2479.*

*Sprenger, M., S. Schemm, R. Oechslin, and J. Jenkner, 2017: Nowcasting Foehn Wind Events Using the AdaBoost Machine Learning Algorithm. Weather Forecast., **32**,* *1079–1099, doi:10.1175/WAF-D-16-0208.1. http://journals.ametsoc.org/doi/10.1175/WAF-D-16-0208.1.*

*Stockdale, T. N., F. Molteni, and L. Ferranti, 2015: Atmospheric initial conditions and the predictability of the Arctic Oscillation. Geophys. Res. Lett., **42**, 1173–1179, doi:10.1002/2014GL062681.*

*Tashman, L. J., 2000: Out-of-sample tests of forecasting accuracy: An analysis and review. Int. J. Forecast., **16**, 437–450, doi:10.1016/S0169-2070(00)00065-0.*

*Tauser, J., and R. Cajka, 2014: Weather derivatives and hedging the weather risks. Agric. Econ. (Czech Republic), **60**, 309–313, doi:https://doi.org/10.17221/11/2014-AGRICECON.*

*Toniazzo, T., and A. A. Scaife, 2006: The influence of ENSO on winter North Atlantic climate. Geophys. Res. Lett., **33**, 1–5, doi:10.1029/2006GL027881.*

*Totz, S., E. Tziperman, D. Coumou, K. Pfeiffer, and J. Cohen, 2017: Winter Precipitation Forecast in the European and Mediterranean Regions Using Cluster Analysis. Geophys. Res. Lett., **44**, 12,418-12,426, doi:10.1002/2017GL075674.*

*Ukkonen, P., A. Manzato, and A. Mäkelä, 2017: Evaluation of thunderstorm predictors for Finland using reanalyses and neural networks. J. Appl. Meteorol. Climatol., **56**, 2335–2352, doi:10.1175/JAMC-D-16-0361.1.*

*van den Hurk, B., F. Doblas-Reyes, G. Balsamo, R. D. Koster, S. I. Seneviratne, and H. Camargo, 2012: Soil moisture effects on seasonal temperature and precipitation forecast scores in Europe. Clim. Dyn., **38**, 349–362, doi:10.1007/s00382-010-0956-2.*

Vihma, T., B. Cheng, and P. Uotila, 2014: Linkages between Arctic sea ice cover, *large-scale atmospheric circulation, and weather and ice conditions in the Gulf of Bothnia, Baltic Sea. Adv. Polar Sci., **25**, 289–299, doi:10.13679/j.advps.2014.4.00289.*

*Vihma, T., 2014: Effects of Arctic Sea Ice Decline on Weather and Climate: A Review. Surv. Geophys., **35**, 1175–1214, doi:10.1007/s10712-014-9284-0.*

Walsh, J. E., W. L. Chapman, and F. Fetterer, 2016: Gridded Monthly Sea Ice Extent and Concentration, 1850 Onward, Version 1. NSIDC: National Snow and Ice Data Center, Boulder, Colorado USA, https://doi.org/10.7265/N5833PZ5. Accessed 01 Jan 2019.

*Wang, L., M. Ting, and P. J. Kushner, 2017: A robust empirical seasonal prediction of winter NAO and surface climate. Sci. Rep., **7**, 1–9, doi:10.1038/s41598-017-00353-y. http://dx.doi.org/10.1038/s41598-017-00353-y.*

*Wegmann, M., Y. Orsolini, E. Dutra, O. Bulygina, A. Sterin, and S. Brönnimann, 2017: Eurasian snow depth in long-term climate reanalyses. Cryosphere, **11**, 923–935, doi:10.5194/tc-11-923-2017. www.the-cryosphere.net/11/923/2017/.*

*Weisheimer, A., and T. N. Palmer, 2014: On the reliability of seasonal climate forecasts. J. R. Soc. Interface, **11**, 20131162–20131162, doi:10.1098/rsif.2013.1162. http://rsif.royalsocietypublishing.org/cgi/doi/10.1098/rsif.2013.1162.*

*Weisheimer, A., N. Schaller, C. O'Reilly, D. A. MacLeod, and T. Palmer, 2017: Atmospheric seasonal forecasts of the twentieth century: multi-decadal variability in predictive skill of the winter North Atlantic Oscillation (NAO) and their potential value for extreme event attribution. Q. J. R. Meteorol. Soc., **143**, 917–926, doi:10.1002/qj.2976.*

Wilks, D. S., 2011: Statistical Methods in the Atmospheric Sciences. 3rd ed. Academic Press, Oxford, 676 pp.

Figures

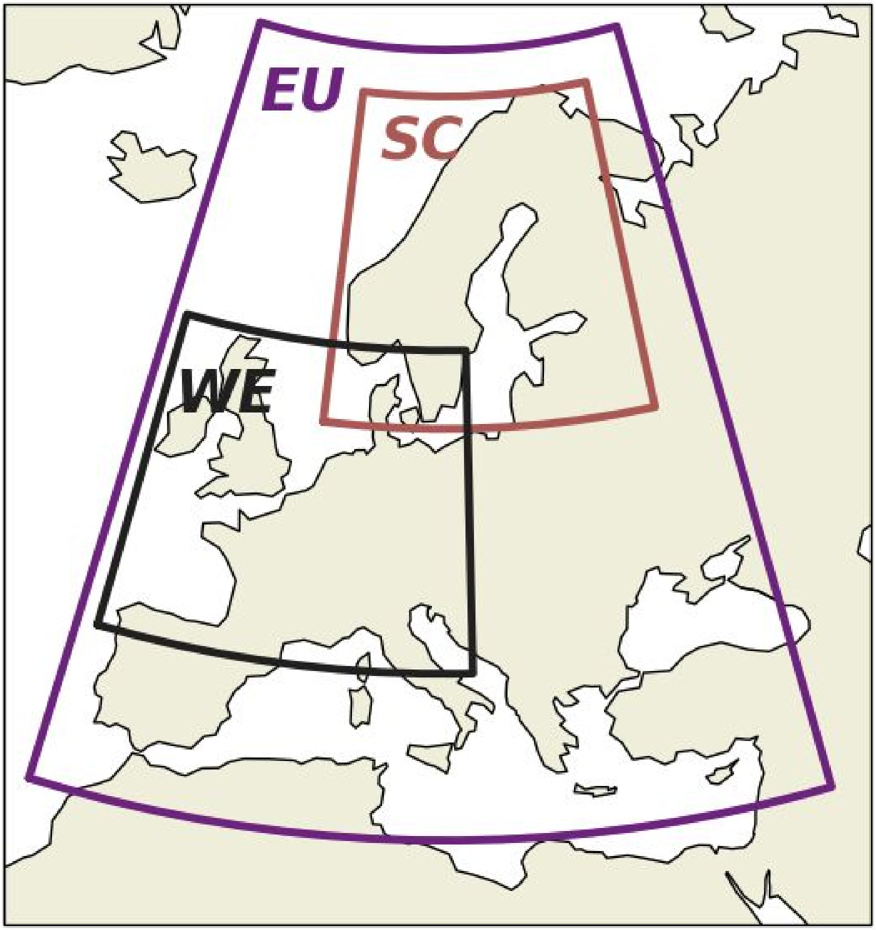


Figure 1. Target domains of this study: Scandinavia (SC; 55°N–71°N, 4°E–34°E), western Europe (WS; 42°N–59°N, 10°W–17°E), and the whole Europe (EU; 33°N–73°N, 12°W–40°E).

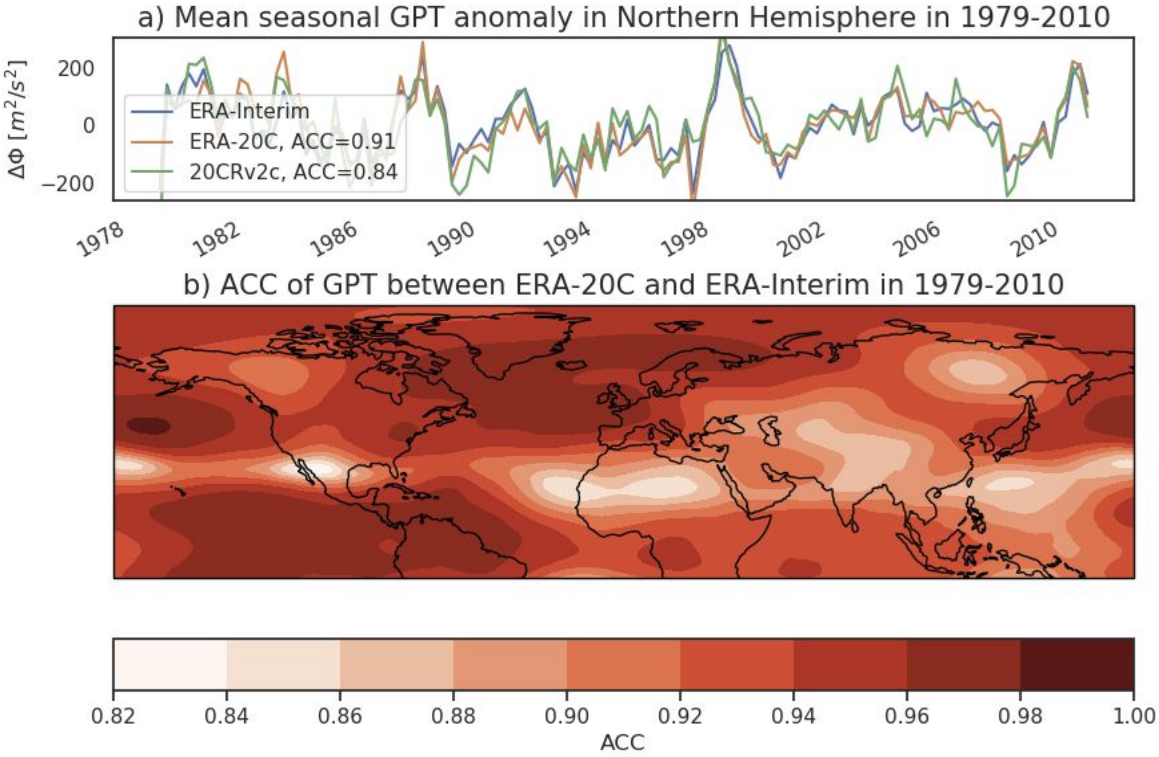


Figure 2. Top: seasonal time series and anomaly correlation coefficient of the mean geopotential anomaly at the 150 hPa level, calculated between the ERA-Interim reanalysis and the reanalyses of this study, ERA-20C and 20CRv2c, in 1979–2010. Bottom: spatial distribution of the anomaly correlation coefficient of the geopotential between ERA-Interim and ERA-20C for the same period.

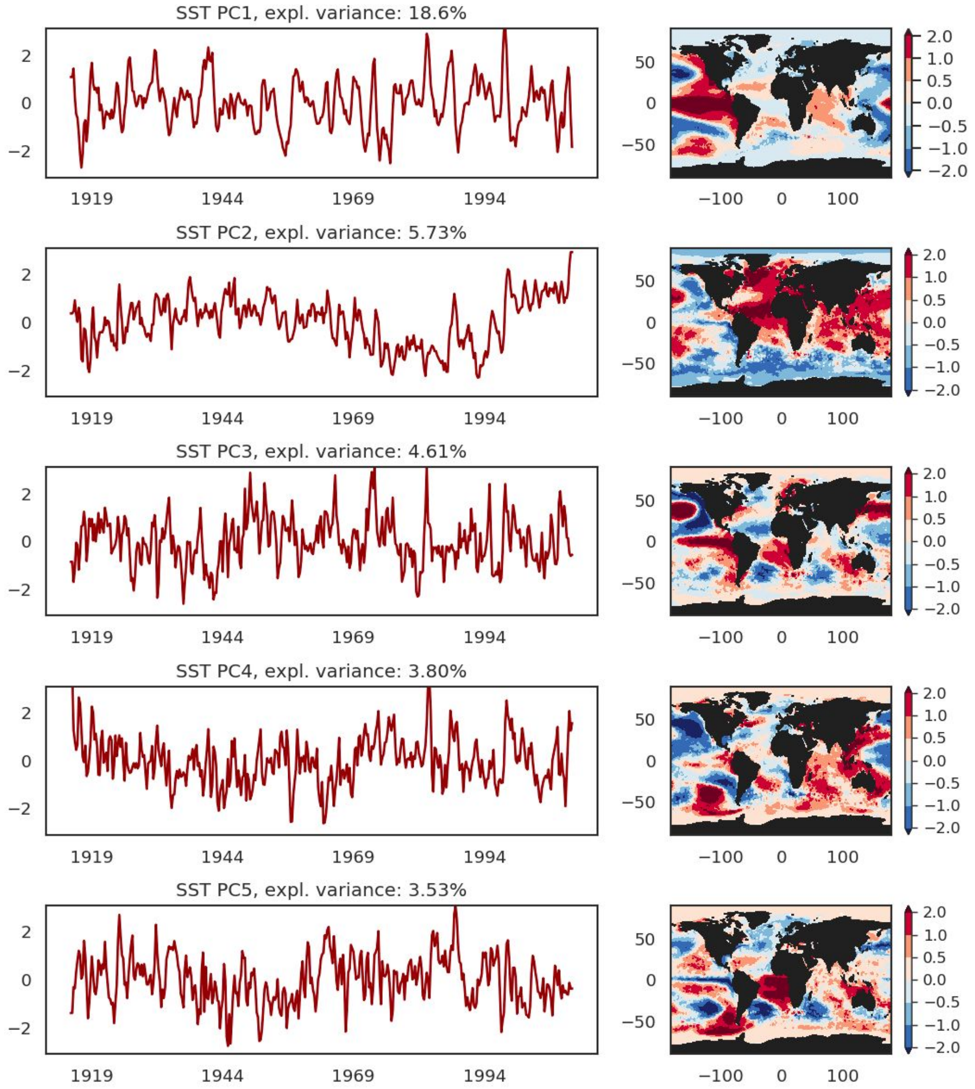


Figure 3. The five N(0,1) -normalized leading PCs of the global sea surface temperature in 1915–2010 (left) and their corresponding, N(0,1) -normalized empirical orthogonal functions (right) in ERA-20C. Explained variance proportions are shown in titles.

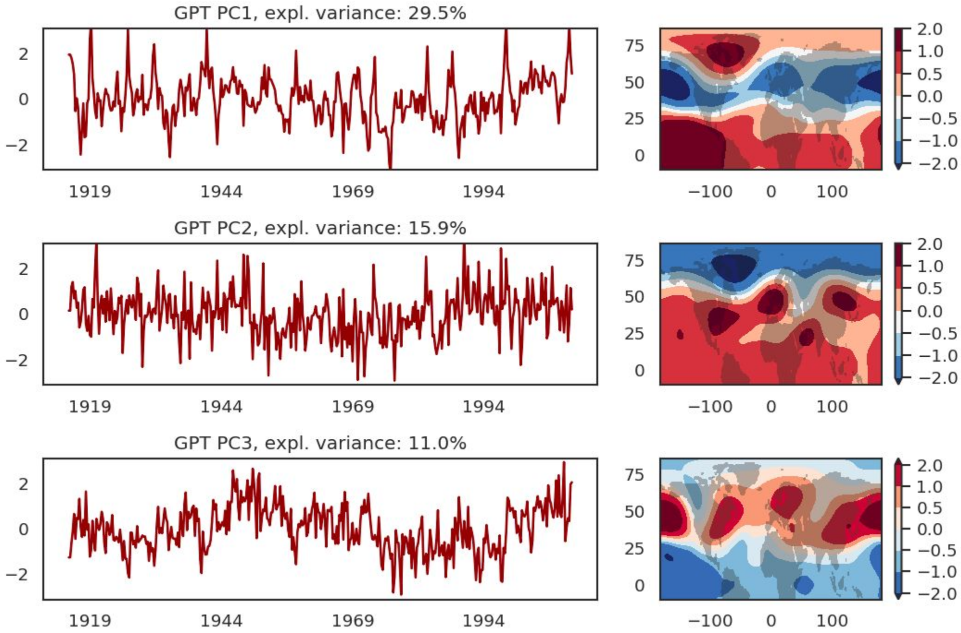


Figure 4. As in Figure 3, but for the geopotential at the 150 hPa level over the Northern Hemisphere.

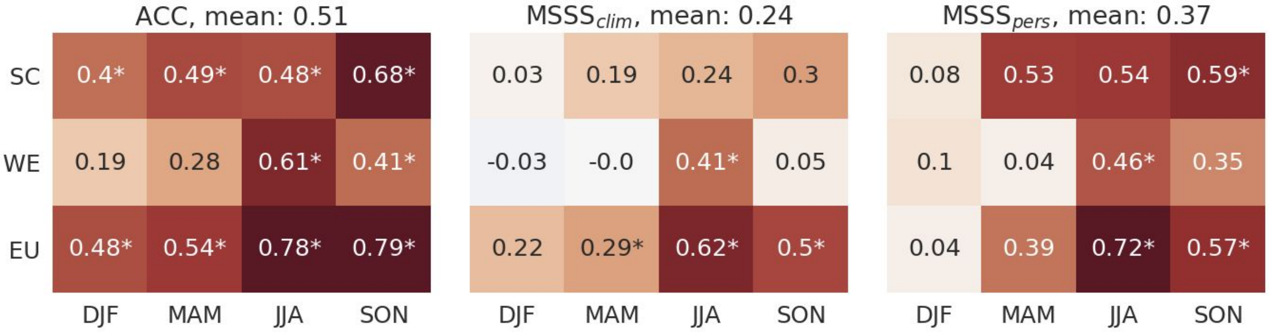
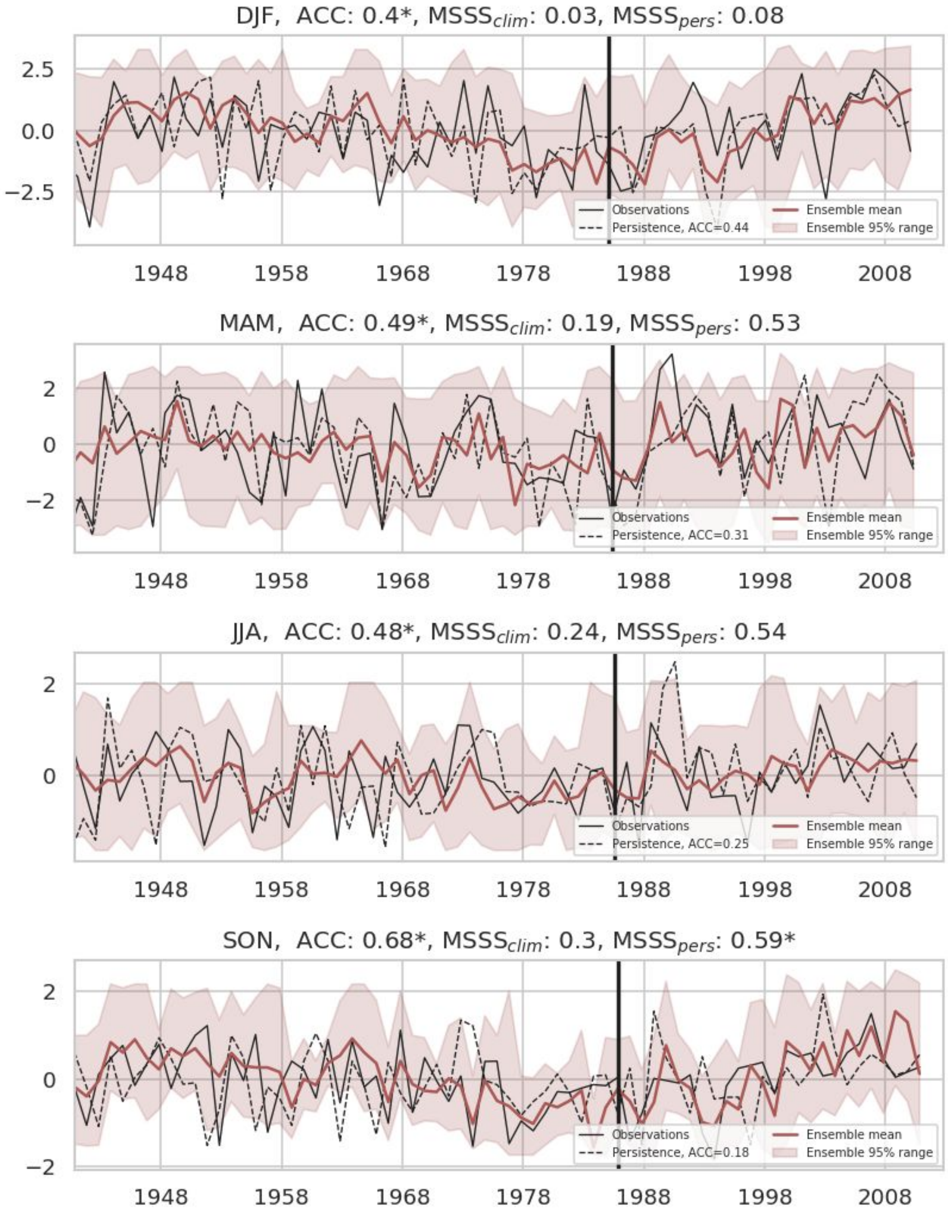


Figure 5. Validation scores in different seasons and target domains, calculated from the independent test years, using the ensemble mean as the best-guess deterministic forecast. Mean values shown in titles. Moving-block bootstrap with 5 years wide block and 10^4 samples was used to estimate the statistical significance (two-sided test, $p < 0.05$, shown with asterisks; Wilks 2011, section 5.3). Colors are used to highlight the validation score values: red tones for positive and blue for negative values.


DJF, ACC: 0.4*, MSSS_{clim}: 0.03, MSSS_{pers}: 0.08

MAM, ACC: 0.49*, MSSS_{clim}: 0.19, MSSS_{pers}: 0.53

JJA, ACC: 0.48*, MSSS_{clim}: 0.24, MSSS_{pers}: 0.54

SON, ACC: 0.68*, MSSS_{clim}: 0.3, MSSS_{pers}: 0.59*

Figure 6. Anomalies of seasonal temperatures in SC *according to observations (solid black), to the persistence forecast (dashed black), and to the ensemble forecast system (ensemble mean is denoted with dark red lines, and ensemble spread with light red shading). Black vertical line denotes the boundary between the training/validation years (1915–1985; beginning excluded from figure for clarity) and independent test years (1986–2010). Skill scores and their significances are defined as in Figure 5.*

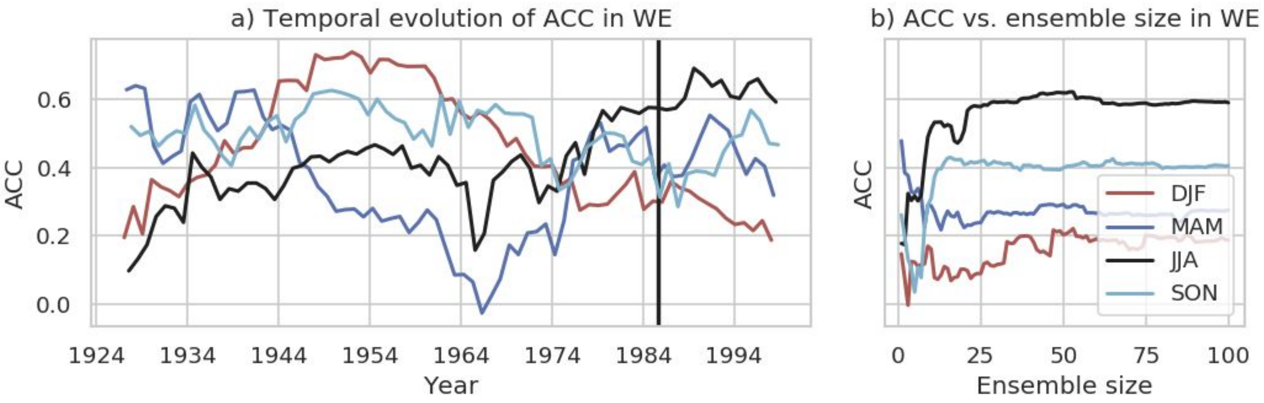


Figure 7. Left: evolution of the ACC skill of temperature forecasts over all years of the study in the WE domain. A moving, 25 years wide analysis window was applied. The middle point of the window is given on the x axis. The black vertical line denotes the boundary between the training/validation years (1915–1985) and independent test years (1986–2010). Right: dependency of the ACC skill on the ensemble size in the test years in the WE domain.

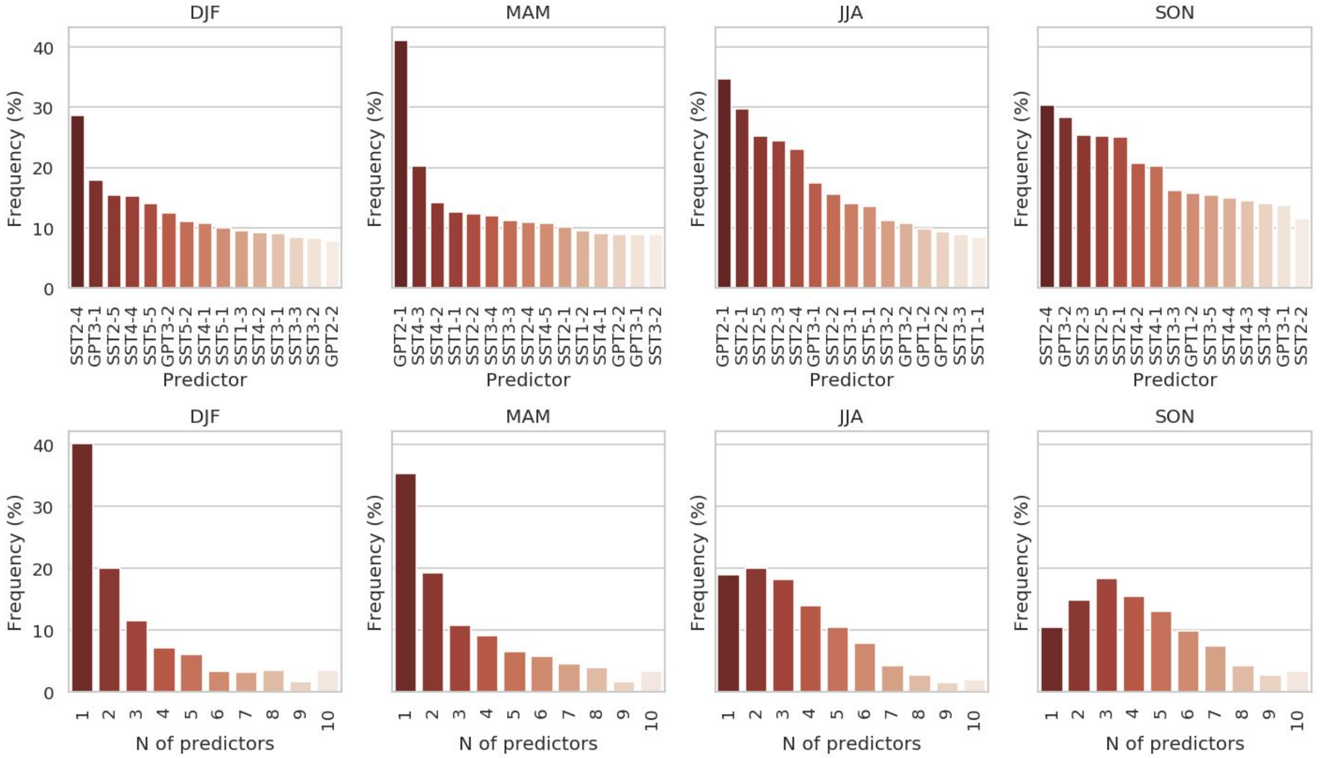


Figure 8. The occurrence of individual predictors (top row; 15 most frequent predictors shown) and the distribution of number of predictors (bottom row) in the LASSO models in the largest EU domain. The number before the dash sign indicates the principal component, and the number after the dash the number of lags of the predictor.