Tellus SERIES A DYANAMIC METEOROLOGY AND OCEANOGRAPHY
PUBLISHED BY THE INTERNATIONAL METEOROLOGICAL INSTITUTE IN STOCKHOLM

Skill and added value of the MiKlip regional decadal prediction system for temperature over Europe

By HENDRIK FELDMANN[1]*, JOAQUIM G. PINTO[1], NATALIE LAUBE[1], MARIANNE UHLIG[1,2], JULIA MOEMKEN[1], ALEXANDER PASTERNACK[3], BARBARA FRüH[4], HOLGER POHLMANN[5], and CHRISTOPH KOTTMEIER[1], [1]*Institute for Meteorology and Climate Research (IMK-TRO), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany;* [2]*NZ Climate Change Institute, Victoria University of Wellington, Wellington, New Zealand;* [3]*Institute for Meteorology, Freie Universität Berlin, Berlin, Germany;* [4]*Deutscher Wetterdienst (DWD), Offenbach, Germany;* [5]*Max-Planck-Institute for Meteorology, Hamburg, Germany*

(Manuscript received 14 November 2018; in final form 8 May 2019)

ABSTRACT

In recent years, several decadal prediction systems have been developed to provide multi-year predictions of the climate for the next 5–10 years. On the global scale, high decadal predictability has been identified for the North Atlantic sector, often extending over Europe. The first full regional hindcast ensemble, derived from dynamical downscaling, was produced within the German MiKlip project ('decadal predictions'). The ensemble features annual starting dates from 1960 to 2017, with 10 decadal hindcasts per starting year. The global component of the prediction system uses the MPI-ESM-LR and the downscaling is performed with the regional climate model COSMO-CLM (CCLM). The present study focusses on a range of aspects dealing with the skill and added value of regional decadal temperature predictions over Europe. The results substantiate the added value of the regional hindcasts compared to the forcing global model as well as to uninitialized simulations. The results show that the hindcasts are skilful both for annual and seasonal means, and that the scores are comparable for different observational reference data sets. The predictive skill increases from earlier to more recent start-years. A recalibration of the simulation data generally improves the skill further, which can also be transferred to more user-relevant variables and extreme values like daily maximum temperatures and heating degree-days. These results provide evidence of the potential for the regional climate predictions to provide valuable climate information on the decadal time-scale to users.

Keywords: regional decadal predictions, added value of downscaling, forecast recalibration, MiKlip

1. Introduction

Decadal prediction aims to forecast multi-year climate trends and anomalies. Such attempts to predict the climate a few years ahead – far beyond the limits of weather prediction – are a quite new endeavour. In fact, the pioneering papers presenting actual decadal hindcasts appeared in and after 2007 (Smith et al., 2007; Keenlyside et al., 2008; Pohlmann et al., 2009; Smith et al., 2013). Decadal prediction gained a lot of interest in the last years, for instance as a branch of the CMIP5 simulation plan (Meehl et al., 2009). Meanwhile, several large hindcast ensembles from different coupled global climate or earth system models are available. This enables the assessment of the general characteristics of the predictive potential and the skill verification on decadal time-scales (Kim et al., 2012; Doblas-Reyes et al., 2013; Bellucci et al., 2015).

It has been established that the highest predictive skill can be found generally for some oceanic regions, particularly for the northern Atlantic. The source of the predictability there is connected to long-term variations of the Atlantic Multidecadal Oscillation (AMO, Sutton and Hodson, 2005; Latif and Keenlyside, 2011). Ding et al. (2016) estimated that the predictability limit for the AMO is up to 11 years. Skill and an added value of initialization were found for multi-year annual sea surface temperature for this region (Kim et al., 2012; Doblas-Reyes et al., 2013; Bellucci et al., 2015). This skill often

*Corresponding author. e-mail: hendrik.feldmann@kit.edu
Supplemental data for this article is available online at https://doi.org/10.1080/16000870.2019.1618678

does not extend far over continental areas (Hermanson et al., 2014). Nevertheless, Europe has a strong potential for predictive skill over land, as it is located downstream of the North Atlantic hot spot of decadal predictability. For instance, Müller et al. (2012) found a predictive skill for multi-year seasonal mean temperature over Europe.

The German research program MiKlip ('Mittelfristige Klimaprognosen', i.e. medium-term climate predictions, Marotzke et al., 2016) is designed to provide a decadal prediction system. It generates, improves and analyses decadal prediction based on the earth system model MPI-ESM. Furthermore, it encompasses modules to improve the initialization, the treatment of relevant processes on the decadal scale and to evaluate the decadal hindcasts. In addition, MiKlip includes a regionalization module, which provides for the first time systematic efforts for dynamical downscaling of decadal predictions, with the regional focus on Europe. The goal of this module is to estimate the predictive skill and added value of high-resolution decadal predictions at the regional scale. First analyses of the regionalized MiKlip hindcasts by Mieruch et al. (2014), Reyers et al. (2019) and Moemken et al. (2016) indicated skill and added value of downscaling and initialization for parts of Europe. In particular, they could show that the dynamical downscaling may reduce the bias and increase the reliability of the hindcasts. Compared to the first two of these previous works, the present study relies on a hindcast ensemble with a larger sample size and annual starting dates from 1960 to 2017. This increased sample size allows for a more robust assessment of the skill and added value of decadal predictions and to investigate several aspects of regional decadal predictions, which have not been analysed in previous efforts, for instance the lead- and averaging time dependence or the temporal evolution of the predictive skill.

Goddard et al. (2013) proposed a framework for a standard verification and analysis for decadal predictions, adapted from weather or seasonal prediction. They propose to analyse lead-years 1, 2–5, 6–9 and a long averaging period years 2–9, since the main sources of the decadal predictability are slow processes mainly arising from the ocean circulation and the multiyear response to natural or anthropogenic forcing. In addition, they propose a spatial averaging over (sub-)continental regions. On the other hand, users typically desire a high temporal and spatial resolution of climate information, but this work will assess these recommendations and test them to examine the appropriate temporal and spatial scales for regional decadal predictions for Europe.

When trying to determine the capabilities of decadal prediction, several aspects have to be considered. First, skill scores are derived from a set of hindcasts, to estimate the general quality of the forecasts, which is giving a measure of what can be expected from the prediction system in the future. The second aspect is to demonstrate that the effort is worthwhile. This search for an added value has different aspects. As usual, the forecasts should be better than climatology. For temperature, the climate trend is a major source of the potential predictability. This trend should already be represented by the so-called 'historical' climate projections. These simulations consider the changes in external radiative climate forcing (e.g. the changes in the greenhouse gas concentrations), but have no information on the phase of the natural variability patterns. Nevertheless, the historical simulations also consider radiative effects from volcanic eruptions and the solar cycle, which can have an influence on climate variability. The decadal hindcast are initialized to match the state of the climate system at the starting date (e.g. SSTs) and is thus expected to capture to some extend the phases of the decadal scale internal climate variations. The added value of initialization is quantified by comparing the initialized hindcasts to the un-initialized climate projections with respect to the representation of these phases. As a third aspect, the potential added value with respect to the lower resolution global predictions in the target region has to be considered.

This work focusses on the characterization of skill and added value for near-surface temperature (tas hereafter) and temperature related variables over Europe. This includes the potential of post-processing to improve the skill. This study characterizes several aspects of regional decadal predictions for Europe, and poses the following research questions:

- Do regional decadal predictions provide skill and added value for distinct regions over Europe?
- How robust are the skill estimates? This includes the consideration of the observation uncertainty.
- How does the skill of the predictions change over time? This includes the lead- and averaging-time as well as the hindcast period.
- Can the skill scores be improved by recalibration?

The paper is organized as follows: After this introduction (Section 1), the decadal MiKlip hindcasts are described. Section 3 focusses on the skill metrics applied in this paper. The results are presented in Section 4 and Section 5 contains the summary and conclusions.

2. The MiKlip decadal hindcast ensemble

The global model of the MiKlip decadal prediction system is the MPI-ESM with the atmospheric component ECHAM6 (Stevens et al., 2013) and the ocean model MPI-OM (Jungclaus et al., 2013). Several generations of

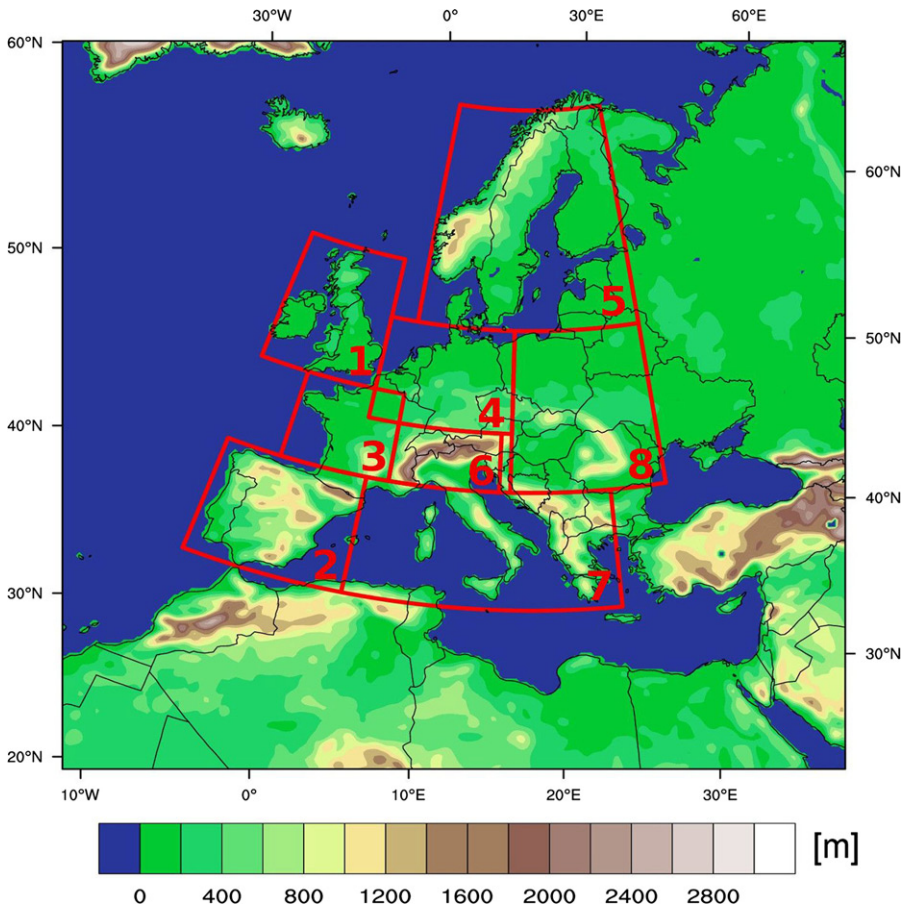

Fig. 1. Model domain with topography for the EUR-44 grid and sub-regions – 1: British Isles (BI), 2: Iberian Peninsula (IP), 3: France (FR), 4: Mid-Europe (ME), 5: Scandinavia (SC), 6: Alps (AL), 7: Mediterranean (MD) and 8: Eastern Europe (EA).

decadal hindcast experiments were performed in MiKlip with different resolutions, initialization and ensemble generation strategies (Matei et al., 2012; Marotzke et al., 2016; Müller et al., 2018) for the period after 1960. This paper focusses on the second generation (named baseline1 or b1 hereafter). For the regionalization the low resolution version MPI-ESM-LR was applied as initial- and boundary-forcing, which has a resolution of 1.875° (T63) with 40 vertical layers (Müller et al., 2012; Pohlmann et al., 2013).

The global hindcast ensemble consists of 10 decadal (10-year) simulations for each starting year. For each year after 1960, a new 10-member set is initialized at the first of January of the following year (start date decadal1960: 1st January 1961). The initial ensemble spread for the global ensemble was generated using a one-day time-lag approach (Müller et al., 2012; Romanova et al., 2018). Details regarding the anomaly initialization of the ensemble generations with observations can be found in Matei et al. (2012). An ensemble of 10 CMIP5 historical simulations was used as a reference to estimate the added value of the initialization. This un-initialized ensemble uses otherwise the same MPI-ESM model version and setup as the global hindcasts.

A dynamical downscaling was performed with the regional climate model (RCM) COSMO-CLM (CCLM; Doms and Schättler, 2002; Rockel et al., 2008) for Europe over the EURO-CORDEX domain at 0.44° (∼50 km) resolution on a rotated latitude–longitude grid. The setup of the models is conforming to the setup used for EURO-CORDEX (Jacob et al., 2014). The initial soil conditions for the regional hindcast simulations were taken from a long-term ERA40/ERAInterim driven CCLM simulation with the same setup (Khodayar et al., 2015). Figure 1 shows the model domain including the boundary zone and the sub-regions used in the analysis.

The downscaling was applied to all members of the MPI-ESM-LR baseline1 hindcast for annual starting dates from 1960 to 2017. Therefore, the regional as well as the global ensemble consists of 5800 simulation years. This provides a much larger sample size compared to previous studies of regional decadal predictions (Mieruch

et al., 2014; Reyers et al., 2019). This larger sample size here allows for more robust skill estimates in this work, compared to the aforementioned studies.

3. Skill metrics

The verification methods within this paper follow the strategy depicted by Goddard et al. (2013). Moreover, some of their recommendations are tested here. This includes, for instance, that averaging over large areas are necessary to determine robust skill estimates and the use of 4-year averages.

The quality of a forecast encompasses several aspects (Wilks, 2011), which have to be determined using objective metrics. These are:

- Accuracy, which indicates the correspondence between the forecasts and the observed events.
- Reliability, which is the statistical consistency between the predicted probabilities and the subsequent observations.

Most commonly used as accuracy metric, the Anomaly Correlation Coefficient (ACC) will be applied to the MiKlip hindcasts to measure the overall skill of a forecast. The ACC is a measure of association that compares anomaly values of forecast and observation in time. Anomalies in this case are the values derived by subtracting the climatological mean (Wilks, 2011). There are two forms of the ACC. In the following, the un-centred anomaly correlation will be used which is equivalent to the Pearson correlation coefficient. A perfect correlation between forecast and observation results in an ACC of 1, anti-correlation in −1. An ACC of zero means no relation at all.

The Mean Squared Error Skill Score (MSESS) is the second metric used here to describe overall skill. The MSESS does not only reflect the association between two time series as does the correlation, moreover it takes unconditional and conditional bias into account, too. Furthermore, it allows determining if the forecast is superior to a reference. The MSESS is defined as follows:

$$\mathrm{MSESS} = 1 - \frac{\mathrm{MSE}_{\mathrm{Fcst}}}{\mathrm{MSE}_{\mathrm{Ref}}} \text{ with } \mathrm{MSE} = \frac{1}{N}\sum_{n}^{N}\left(\overline{X_i} - O_i\right)^2 \quad (1)$$

MSE is the mean square error, with $\overline{\mathrm{X}}$ as the ensemble mean of the forecast and O the observed value. For both data sets, anomalies are used. 'Fcst' denotes the selected forecast and 'Ref' the respective reference forecast. A perfect agreement between forecast and observation means a skill score of 1. Other than for the ACC, MSESS has no lower bound.

If the reference is the climatological mean ($\overline{O}$), than MSESS can be decomposed into the following terms (Murphy, 1988):

$$\mathrm{MSESS}\left(X, \overline{O},\ O\right) = r^2_{\mathrm{Fcst,o}} - \left[r_{\mathrm{Fcst,o}} - \left(\frac{S_{\mathrm{Fcst}}}{S_{\mathrm{O}}}\right)\right]^2 - \left[\frac{\overline{X} - \overline{O}}{S_O}\right]^2 \quad (2)$$

$r_{\mathrm{Fcst,o}}$ denotes the correlation between the forecast and the observation and $\left(\frac{S_{\mathrm{Fcst}}}{S_O}\right)$ the ratio of the standard deviations.

The first term is the correlation between the forecast and the observation and the second term is the conditional bias (CB), which is a measure of the reliability of the forecast. The third term represents the unconditional bias, which can be used to analyse anomalies.

An often-applied measure of the reliability is the Continuous Ranked Probability Skill Score (CRPSS):

$$\mathrm{CRPSS} = 1 - \frac{\mathrm{CRPS}_{\mathrm{Fcst}}}{\mathrm{CRPS}_{\mathrm{Ref}}} \text{ with: } \mathrm{CRPS} = \int_{-\infty}^{\infty}\left[F(y) - F_{\mathrm{o}}(y)\right]^2 \mathrm{d}y \quad (3)$$

As for any skill score, the perfect score is 1. A skill of 0 or lower indicates no skill in comparison to the reference.

Positive skill scores of MSESS and CRPSS are used as a quantitative measure of the added value of a hindcast compared to a reference. The added value for the correlation is determined by the differences of the correlation coefficients

$$\mathrm{ACC}_{\mathrm{AV}} = r_{\mathrm{Fcst}} - r_{\mathrm{Ref}} \quad (4)$$

For the conditional bias, the added value is defined as the difference of the absolute values of CB:

$$\mathrm{CB}_{\mathrm{AV}} = \left|r_{\mathrm{Ref},o} - \left(\frac{S_{\mathrm{Ref}}}{S_O}\right)\right| - \left|r_{\mathrm{Fcst},o} - \left(\frac{S_{\mathrm{Fcst}}}{S_O}\right)\right| \quad (5)$$

The skill assessment uses the forecast verification suite described by Kadow et al. (2016) and the so-called MurCSS plugin (Illing et al., 2014) of the MiKlip central evaluation system CES. The MurCSS tool includes a lead-time dependent calculation of anomalies in line with ICPO (2011). Throughout the paper, only anomalies are used, to avoid a sensitivity of the metrics with respect to potential biases in the climatological mean of the ensembles. The significance of the results at the 5% level was determined with 500 times bootstrapping, by calculating the scores with a random re-sampling of the time-series with replacement. Auto-correlation is taken into account.

4. Characterization of the regional ensemble

4.1. Observational uncertainty for the skill assessment of regional decadal predictions

A high quality observational data set is crucial to assess the skill of climate predictions. Ideally, the data set should cover the full hindcast period and the whole region of interest. Furthermore, it should be temporarily and spatially homogenous. These goals are difficult to meet over time-scales of more than 50 years. However,

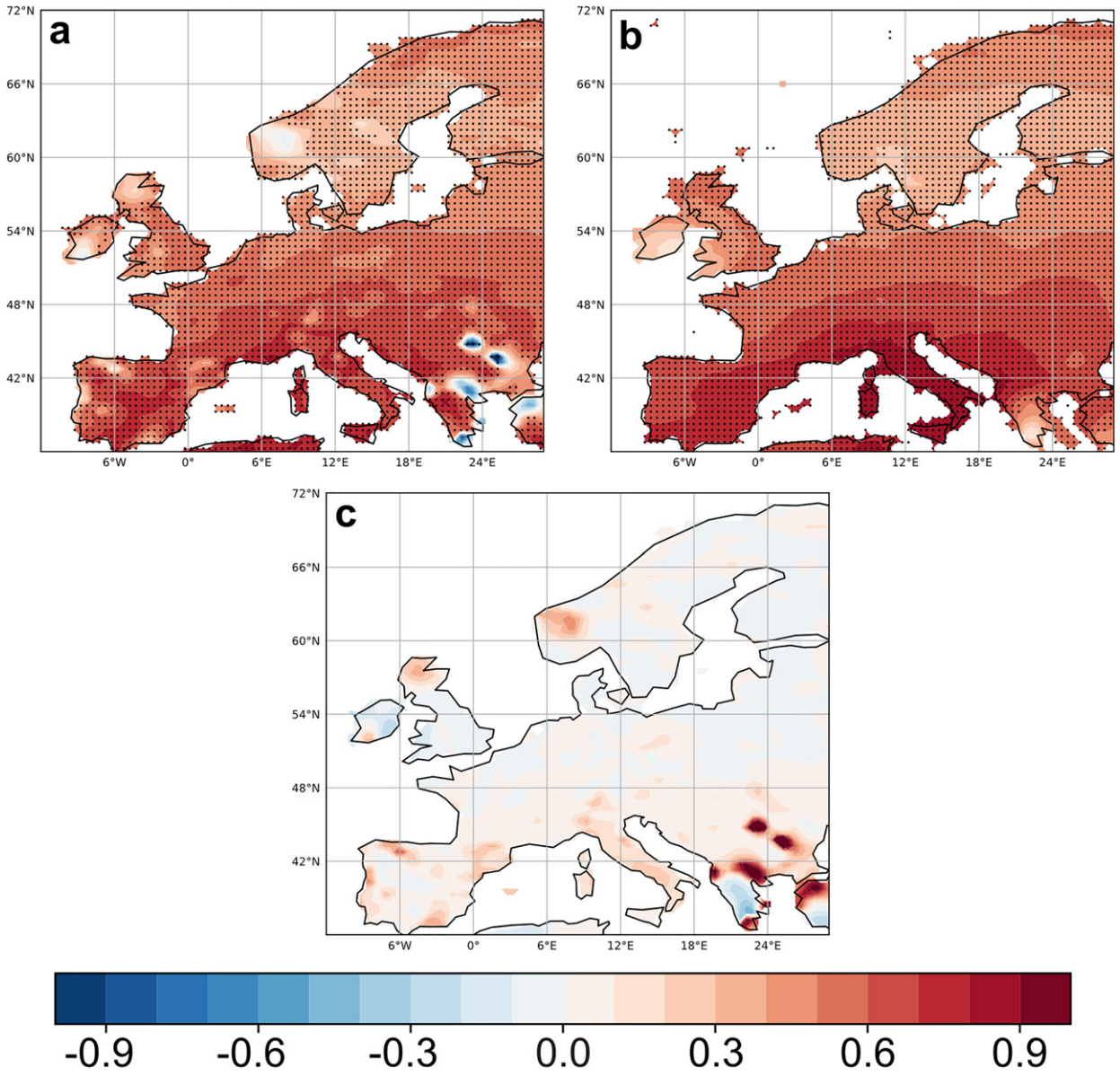

Fig. 2. MSESS of CCLM hindcasts (average over lead-years 2–5) w.r.t. climatology derived from (a) E-OBS and (b) CRU TS4.01 for the period 1962–2015. The black dots indicate significant skill at the 95% level. (c) Difference MSESS CRU – MSESS E-OBS.

such long periods are needed to achieve robust skill estimates to encompass the internal climate variability over a couple of decades. For Europe, several gridded observational data sets are available. Here, we use mainly two such data sets: E-OBS (Haylock et al., 2008) which provides daily data of mean, maximum and minimum temperatures, and CRU TS4.01 (Harris et al., 2014), which provides monthly values of these parameters. Both observational references are updated continuously and cover the whole analysis period from 1961 until 2015.

However, the measurements are also subject to uncertainties and errors. For instance, they are not necessarily homogenous – neither in space or time. Only parts of the station observations used to derive these datasets are homogenised by various national meteorological services using different methods (Harris et al., 2014). Moreover, the density, quality and number of stations with sufficiently high data quality varies between the European countries, with good coverage in Central Europe and partly less in areas like Eastern and Southern Europe (e.g. Hofstra et al., 2011; van der Schrier et al., 2013).

To demonstrate the impact of these uncertainties on the evaluation of the decadal forecasts we used both data sets for computing the hindcast skill of the regional

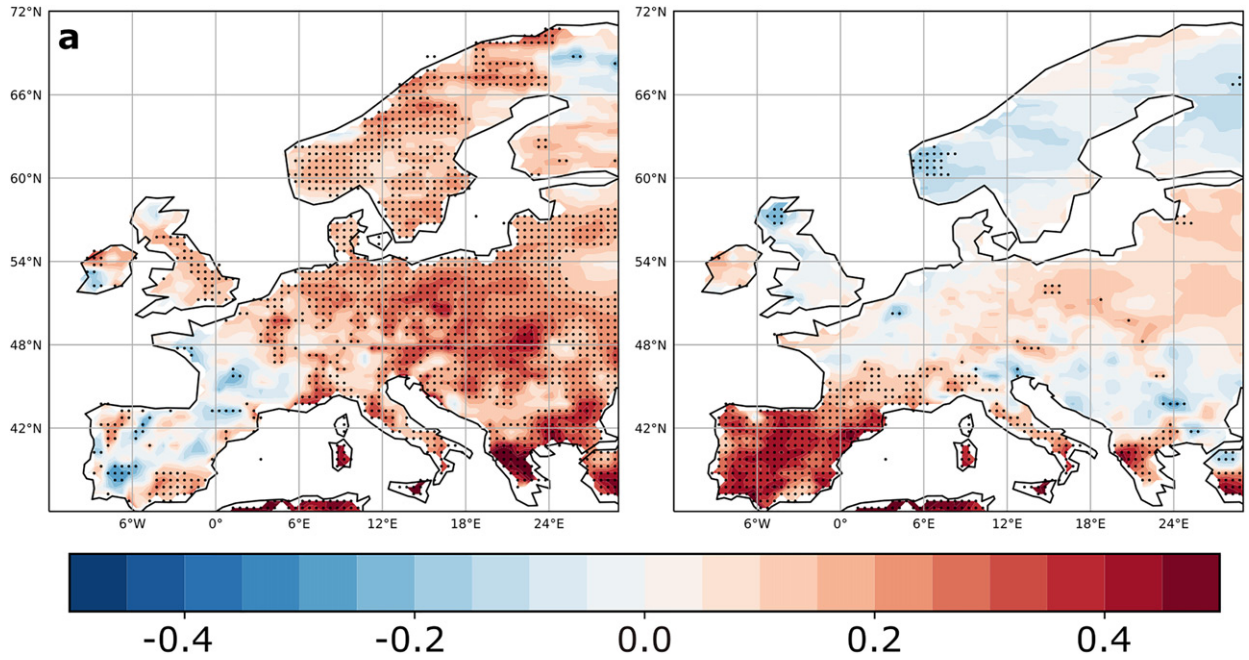

Fig. 3. Skill and added value of downscaling: MSESS near-surface temperature (tas) compared to E-OBS; period 1967–2016, lead-time year 2–5. (a) MSESS CCLM b1 vs. MPI-ESM historical, (b) MSESS CCLM vs. MPI-ESM-LR b1.

decadal ensemble for the temporal mean lead-years 2–5 after initialization. Figure 2 shows the MSESS (supporting information Fig. S1 for correlation) with the climatology as reference and both gridded observational reference data sets. The verification against both references display the same major features: A significant skill over most of Europe with the highest values for Southern Europe and partly lower values in Eastern Europe and Scandinavia. The skill pattern of the CRU data is smoother compared to E-OBS. This might be partly due to a lower number of stations, with less small-scale structures for CRU, which might also not fully be represented by the hindcasts. For some small areas in the Iberian Peninsula, Italy and especially in South-Eastern Europe, there are grid boxes, where the comparison with E-OBS shows extremely negative local skill measures. Such negative values are often very local and do not occur for neighbouring grid boxes or in the skill scores using CRU as reference. This might indicate inhomogeneous data included for these boxes in E-OBS.

The difference of the skill scores between the two reference sets is shown in Fig. 2c. The differences are small over most of the domain (e.g. UK, France and Germany), where there are a large number of observations available. Larger differences occur mainly for Southern Europe and Northern Africa and for South-Eastern Europe. In particular, some areas with negative skill score vanish when using the CRU instead of the E-OBS data as observational reference, namely over Southern Greece, Turkey, Northern Africa and a few grid boxes in South-Eastern Europe. This seems to hint that the very low skill values there compared to neighbouring areas may be related to problems in the E-OBS data and not to a peculiar behaviour of the decadal hindcasts.

The same features are identified for the other skill metrics. The skill scores averaged over the sub-regions (supporting information Fig. S2) do not differ much for areas with good data coverage, namely the British Isles (BI), France (FR), Mid-Europe (ME) and parts of Scandinavia (SC). Larger differences are found for Southern Europe (MD, IP, AL) and especially for the Mediterranean (MD) region, which shows the highest mean skill using CRU as a reference. The skill scores are lower for E-OBS, a fact that can be partially attributed to a few very large negative values. The differences are lower for the correlation (difference <0.05) than for MSESS (skill difference up to 0.15). Therefore, we conclude, that the results are uncertain due to the shortcomings of the observational reference in some regions. Nevertheless, both data sets provide a useful reference for the verification of the regional decadal hindcasts and are used within this work. The qualitative results remain very similar when using either of them (cf. supporting information Fig. S2).

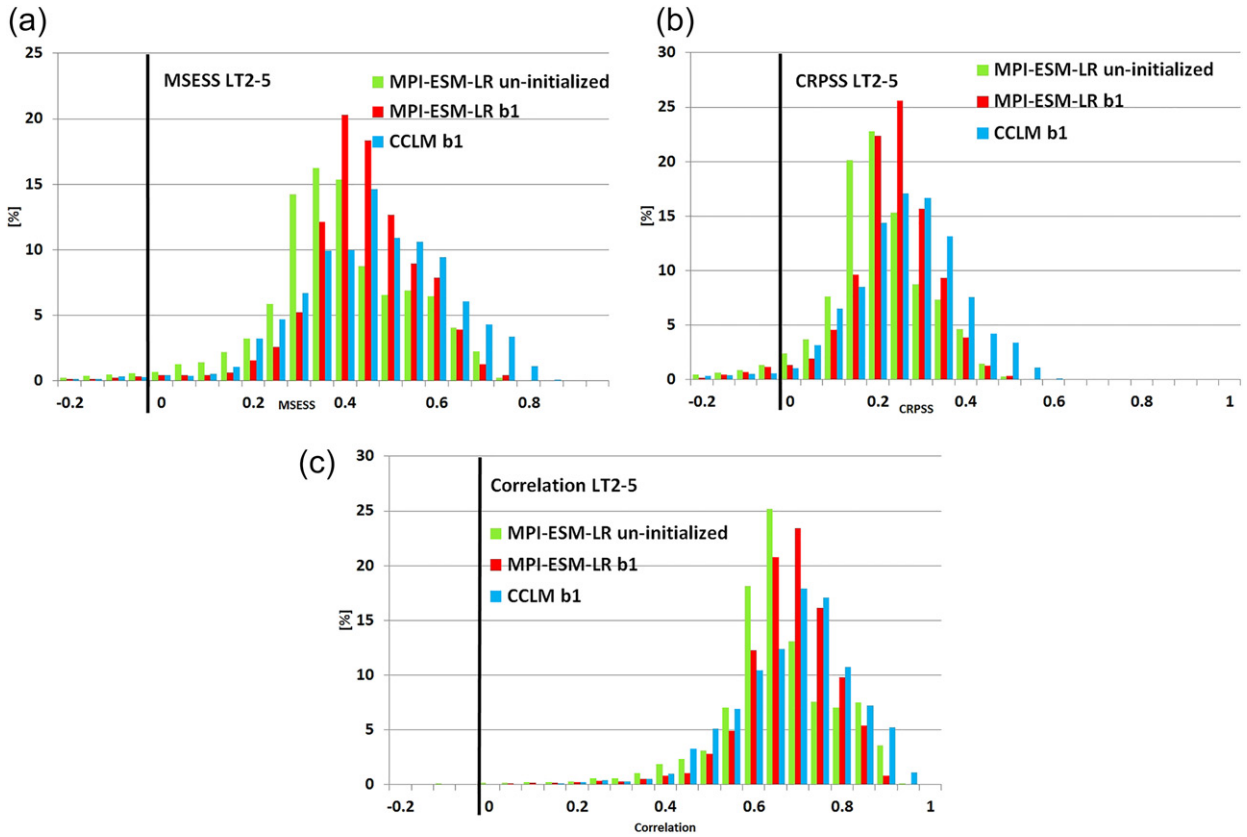

Fig. 4. Skill distribution over Europe for lead-time years 2–5 for CCLM b1 (blue) and MPI-ESM-LR b1 (red) and MPI un-initialized (green). (a) MSESS, (b) CRPSS, (c) correlation (ACC).

4.2. Regional skill and added value of the regional decadal hindcasts

The skill of the hindcast ensembles for temperature is determined using the Mean Square Error Skill Score (MSESS), the Continuous Rank Probability Skill Score (CRPSS) and the correlation (cf. Section 3) to quantify several aspects of the skill. E-OBS is used as observational reference and the verification period is 1967–2015, which is the common period for all 4-year mean lead-time periods (cf. Section 4.3).

The added value of the regional hindcasts is measured by comparing against two reference hindcasts: first, against the un-initialized MPI-ESM-LR historical simulations, to estimate the added value of initialization; second, against the driving MPI-ESM-LR b1 hindcasts, to estimate the added value of downscaling. For each of these hindcast sets, 10 ensemble members are used in the calculations. The mean over lead-years 2–5 from the hindcasts is compared to 4-year means over the same target period for observations and the continuous historical simulations.

The MSESS pattern of the MPI-ESM-LR (not shown) and the regional CCLM hindcasts (c.f. Fig. 2) with the climatology as reference display a similar regional distribution, with positive values over most of Europe and high MSESS scores on the British Isles, France and Central Europe and the highest values around the Mediterranean Sea. MSESS is slightly lower for parts of Scandinavia and Eastern Europe. The added value due to initialization, depicted from the comparison with the un-initialized historical simulations (Fig. 3a), is spread over Europe, with the highest scores for the Eastern Mediterranean region, followed by Eastern, Central Europe, Scandinavia and the UK. These differences are mostly significant. Slightly negative scores are found for parts of Western Europe with the Iberian Peninsula and France. Skill reduction there is mostly not significant, except in a few small areas.

Figure 3b depicts the MSESS of the CCLM hindcasts with the MPI-ESM-LR b1 hindcasts as reference. Positive values indicate an advantage of the regionalization, while negative values indicate the opposite. The downscaling maintains the skill inherited from the global hindcasts, but shows a significantly increased skill up to ca. 0.6 for the Iberian Peninsula and the Mediterranean Region, while a smaller added value can be found for

Table 1. Median and quartiles for the skill metrics (SM): MSESS, CRPSS and ACC compared to the E-OBS observations for regional CCLM b1 ensemble, the initialized MPI-ESM-LR b1 hindcast ensemble and the un-initialized MPI-ESM-LR historical simulations.

Score	Model	Median	SM < 0	0 < SM < 0.25	0.25 < SM < 0.5	0.5 < SM < 0.75	SM > 0.75
MSESS	MPI hist	0.402	3.7%	8.8%	60.4%	26.1%	0.2%
	MPI b1	0.458	2.7%	3.5%	58.6%	34.6%	0.4%
	CCLM b1	0.491	2.4%	5.6%	45.9%	41.3%	4.6%
CRPSS	MPI hist	0.221	5.6%	56.2%	37.5%	0.3%	0.0%
	MPI b1	0.262	4.1%	39.8%	55.8%	0.4%	0.0%
	CCLM b1	0.291	3.1%	33.6%	58.7%	4.6%	0.0%
ACC	MPI hist	0.677	0.2%	1.1%	6.4%	66.5%	25.8%
	MPI b1	0.712	0.1%	0.7%	3.0%	64.2%	32.0%
	CCLM b1	0.727	0.1%	0.4%	5.5%	52.7%	41.3%

parts of Eastern and Central Europe and Ireland. Lower scores of the regional compared to the global hindcasts are found mainly for Scotland and Scandinavia. These negative scores are mostly not significant.

Figure 4a shows the frequency distribution of MSESS for the three ensembles for the European region. The histogram for CCLM b1 is slightly broader than for MPI-ESM-LR b1, but shifted towards higher skill scores for b1. Table 1 confirms these findings, depicting a higher median and more MSESS values in the highest quartiles. This indicates an added value of the downscaling with respect to the accuracy of the hindcasts. In addition, the histogram for the un-initialized ensemble is shifted towards lower values and has a lower median value compared to both initialized ensembles. This further emphasizes an added value of the initialized hindcasts and that the skill for temperature is not only due to externally forced climate change.

The CRPSS (supporting information Fig. S3a, b) measures the reliability, resolution and uncertainty of the hindcasts. The CRPSS scores are positive over most of Europe. The regional distribution is similar to that of the MSESS. An added value of downscaling is found for the Iberian Peninsula, Italy and further parts of the Mediterranean region and Eastern Europe (cf. supporting information Fig. S3b). An added value of the initialization is found from Scandinavia over Central and Eastern Europe towards South-Eastern Europe (supporting information Fig. S3a). A lower reliability was found for the Iberian Peninsula and Western France. The score histogram (Fig. 4b and Table 1) again indicates a significant shift towards higher values of the hindcasts compared to the un-initialized ensemble, as well as a further significant shift of the regional compared to the global hindcasts. These shifts of the distributions are significant at the 95% level according to a *t*-test.

The correlation (ACC, cf. supporting information Fig. S3c, d and Table 1) is high for both hindcast sets, with a median >0.7. Therefore, the range for further improvement is limited. Still there is a significant shift in the frequency distribution for CCLM towards the uppermost quartile (Fig. 4c). The regional distribution of the skill improvement is comparable to the other skill scores, with positive values for the Iberian Peninsula, around the Mediterranean Sea, Ireland and parts of Central and Eastern Europe. On the other hand, the added value of initialization is positive over most of Europe, except Spain and Western France.

4.3. Lead-Time dependence of the decadal hindcast skill

To assess the lead-time dependence, the skill scores were calculated for all 4-year periods from lead-years 1–4 to lead-years 7–10. Each calculation covers the same target period from the first possible 4-year mean value 1967–1970 (lead-years 7–10 of the first hindcast starting year 1960) to 2012–2015 (last full 4-year period of available observations in the dataset). The starting years were shifted accordingly to achieve an identical analysis period.

In the Western part of Europe [Iberian Peninsula (IP), France (FR) and British Isles (BI)], the skill decreases from the first four years of the hindcasts towards intermediate lead-times (cf. Fig. 5). This indicates an impact of the initialization on the skill scores. For the British Isles, the lowest MSESS occurs for lead-time years 3–6. For even longer lead-times, the MSESS shows a slight upward trend. This increasing skill can be attributed to the climate trend rather than the initialization. For Eastern Europe and to a lesser degree the Mediterranean, MSESS is lower in lead-years 1–4 as in lead-years 2–5, which is due to a mean cold bias in year 1 in parts of these regions (which affects the skill there negatively). This effect might be attributed to the soil initialization in those regions (Kothe et al., 2016) and vanishes after year

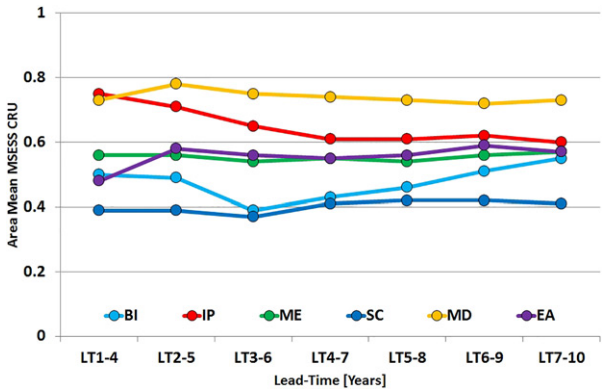

Fig. 5. Lead-time dependence of MSESS for 4-year mean annual temperature in the CCLM hindcast ensemble as average over selected PRUDENCE regions (cf. Fig. 1) for the common verification period 1967–2015. The reference is the climatology of the CRU observations over the period 1967–2015.

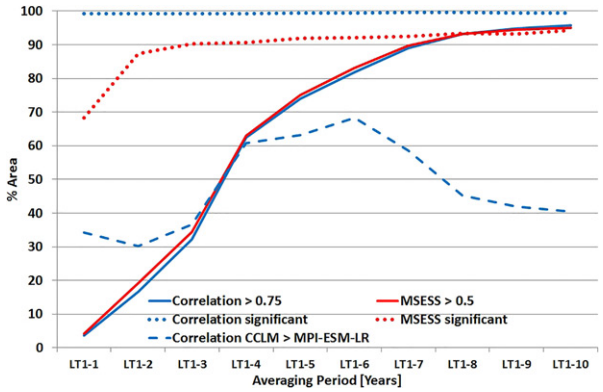

Fig. 6. Area fraction of a considerable skill score in dependence to the averaging period (years) for near surface temperature from CCLM baseline1 ensemble 1961–2015. Anomaly correlation (blue); MSESS (red); area fraction with MSESS >0.5 and ACC >0.75 (solid lines), area fraction with skill significant at the 95% level (dotted), area fraction where skill of CCLM is higher than skill of MPI-ESM-LR (dashed).

1. For the other regions, the skill varies only slightly with the choice of the lead-time window. The level of skill shows a trend from South to North, with the highest MSESS values for the Mediterranean and the Iberian Peninsula (area mean MSESS: 0.6–0.8). For Central and Eastern Europe, the range is ~0.5–0.6. The Northern regions (British Isles and Scandinavia) depict the lowest area mean skill scores (MSESS ~0.4–0.5). The skill decrease originates in winter (c.f. Section 4.6) with its large year-to-year variability.

MSESS for the longest lead-time (year 7–10) shows a common level of about 0.6 for most regions. This indicates the level of skill, which can primarily be attributed to the long-term trend.

4.4. Dependence of skill and added value on the averaging period

To test the effect of temporal averaging, we analysed averaging periods from one year (LT1–1) up to 10 years (LT1–10) all covering the same period 1961–2015. This means that fewer starting years are taken into account for longer averaging periods because it is possible to evaluate LT1–1 for 2015 but not LT1–10 for 2015–2024.

The mean skill scores (averaged over all land grid points) compared to the E-OBS data increases with the length of the averaging period (supporting information Fig. S4). For averaging periods of 4 years and longer the mean skill exceeds 0.5 for MSESS and 0.75 for the correlation. However, the increase in skill is lower for averaging periods longer than four years.

Figure 6 shows the dependence of the skill score from the averaging periods and the area fraction with high skill (MSESS and ACC >0.5). Obviously even for short averaging periods, the skill is significant over most parts of Europe. Almost all land grid points exhibit a significant correlation at the 95% level for all averaging periods. For the MSESS the area fraction is at 68% for year one, 87% for the first two years and beyond 90% for longer periods. The fraction of the area with high skill scores (MSESS >0.5, ACC >0.75) also increases with longer averaging periods. Beyond three-year averages, the skill is high for more than half of Europe, reaching an area fraction of 90% for 7-year averages.

An added value of downscaling indicated by a at least 0.05 higher correlation coefficient is found for averaging periods between four and seven years. For the short averaging periods (e.g. year 1 and years 1–2) the low skill in Eastern Europe in year one reduces the overall fraction of the domain with a higher correlation. For long averaging periods, a lower mean trend of the downscaling ensemble compared to the driving model might be the reason of a reduced added value. These findings are in line with those by Uhlig (2016), who found and optimum skill and added value for averaging periods of 5–7 years.

We conclude that 4-year mean values seem to be a reasonable compromise between the preferred high temporal resolution and a high hindcast skill for near surface temperature. The additional gain of mean skill and for the area fraction of high skill and significance is low, when the averaging period is extended beyond the 7 years.

4.5. Representation of the temporal evolution

The temporal evolution of the 4-year annual mean temperature since 1960 is shown in Fig. 7 for the Mediterranean region as an example, as for this region

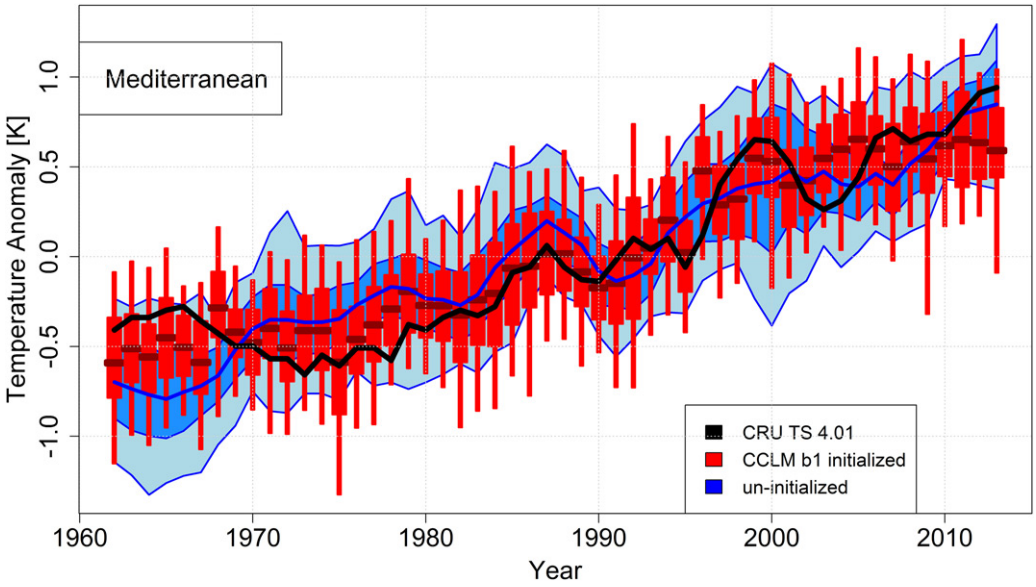

Fig. 7. Temporal evolution of the 4-year mean annual temperatures in the Mediterranean region over the period 1962–2015 from CRU TS4.01 (black), the ensemble of historical simulations (blue), and the CCLM b1 hindcast ensemble (red) for lead-time years 2–5. The light blue area denotes the full range of the historical ensemble; the mid-blue area depicts the inter-quartile range and the dark blue line the ensemble mean. The red lines and boxes indicate the full and interquartile range of the hindcasts, the dark red lines the ensemble mean.

the long-term variability pattern are most discernible. The observations indicate a slight cooling from the end of the 1960s until the mid-1970s. This is followed by a warming trend, which is interrupted by slightly cooler years after the volcanic eruptions of Pinatubo in the year 1991. From the mid-1990s until about the year 2000, a strong warming is observed. All simulated data sets show a similar overall increasing temperature trend. Both ensembles indicate the cooling related to the Pinatubo eruption. This has to be expected, since the external forcing due to volcanoes is applied in both ensembles. In fact, this indicates that even the un-initialized simulations experience some external forcing which affects the climate variability on decadal scales. Their skill in representing the actual temporal evolution is therefore potentially higher than can be expected from the long-term climate trend alone. In other periods with major deviations from the warming trend, for example, the cooling phase from the 1960s to the 1970s or the strong warming in the end-1990s is more closely resembled in the hindcasts compared to the un-initialized time series. This indicates an added value of initialized simulations several years ahead. The temporal evolution for the other regions depict similar but partly less pronounced features (not shown).

The hindcast skill with respect to the climatology or the un-initialized simulations depends on the selected hindcast period. This is due to stronger or weaker signals of the climate variability and trend, but also due to better availability and quality of observational data for initialization and verification in the last decades. To evaluate this dependence for Europe, the full hindcast period has been divided in 20-year segments shifted by 5-years, for example, starting years 1960–1980, 1965–1985 … 1990–2010. For these periods, the correlation with observations has been calculated for the whole domain and (a) the first simulation year (LT1) and (b) for lead-years 2–5 (LT2–5, cf. Fig. 8) separately. The correlation for the early period (1960–1980) and lead-time 2–5 is low. It increases during the later periods, with a slightly reduced skill in the last period. These might point to a better initialization of the relevant processes on a multi-year time-scale for more recent periods. A similar but less pronounced tendency is found for the first lead-year. Comparing the correlation of the initialized against the un-initialized simulations, it becomes obvious that the added value of initialization for lead-year one decreases over time. This may indicate a stronger contribution of the overall warming trend, which is also represented in the un-initialized simulations. On the other hand, the correlation differences increase for the longer lead-time year 2–5, again indicating an improved initialization of the processes relevant for the multi-year variability.

4.6. Seasonal dependence of the skill

To assess the seasonality of the skill pattern, the same analysis as for the annual means has been performed for the 4-year-seasonal means (DJF, MAM, JJA and SON)

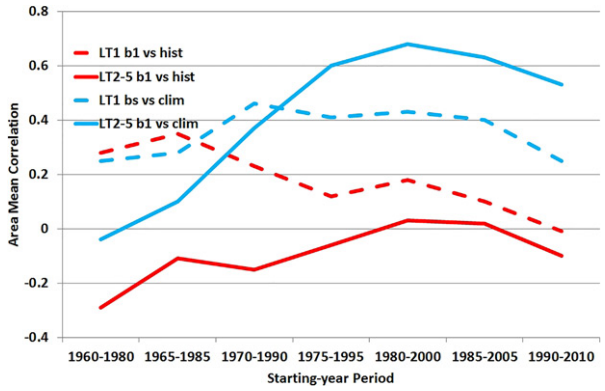

Fig. 8. Ensemble mean correlation of the 4-year mean annual temperature of the regional baseline1 ensemble with the CRU TS4.01 observations for Europe. Anomaly correlation compared to climatology (blue lines) and the un-initialized simulations (red lines); differences between the correlation of the initialized and un-initialized ensembles – mean lead-years 2–5 (solid); lead-year 1 (dashed lines).

as for the annual mean temperature. Figure 9 depicts the correlation pattern for the different seasons and lead-years 2–5. The calculation for DJF starts at month 12 of the first year of the selected period and ends in February in the year after this period.

The skill patterns for the different seasons differ distinctly (Fig. 9). A significant correlation for most of Europe is found for MAM, JJA and SON, displaying comparable area mean skill values. For DJF, the skill is considerably lower, significant only in parts of Europe. Whereas, the skill metrics (Fig. 10) are highest in spring and especially in summer in the southern and eastern parts of Europe, the maximum occurs in spring and autumn in Western and Northern Europe. For the British Isles, France and Scandinavia, the highest values are found in spring. For most regions, the annual skill is higher than for the individual seasons hinting at compensating effects within the seasonal cycle. Exception is the Mediterranean region, where the skill is highest in summer. This high skill for Southern Europe is in line with findings from Müller et al. (2012) and Guemas et al. (2015).

4.7. Recalibration

There is a potential to improve the decadal prediction by post-processing. Within MiKlip Pasternack et al. (2018) developed the Decadal Climate Forecast Recalibration Strategy (DeFoReSt) to recalibrate the raw simulation data against observation. DeFoReSt accounts for a lead- and start-time dependent unconditional bias, conditional bias and ensemble dispersion. This method assumes third- and second-order polynomials to capture lead-year dependent errors and first-order for start-time dependency (linear trend). A cross-validation is included. This method is applied to recalibrate the regional downscaling ensemble with CCLM for the period 1967–2015. The CRU TS4.01 data were used for the recalibration as for the verification.

The recalibration of the CCLM hindcasts for near-surface temperature improves the MSESS compared to the climatology for most of Europe for lead-years 2–5 (Fig. 11a–c). The impact is lowest for the British Isles and Sweden. The improvement is significant for the Southern half of Europe – including the Iberian Peninsula, most of France and parts of Central and Eastern Europe, Italy and the Balkan region. Another region with significant improvement is South-Western Norway. Correlation and conditional bias both contribute to this higher MSESS. ACC increases slightly over most of Europe with significant improvement from Southern to Central Europe and large parts of Scandinavia (Fig. 11d–f). The conditional bias of the recalibrated model is mostly lower than for the raw data except for parts of the British Isles, Southern Scandinavia and the Iberian Peninsula (Fig. 11g–i).

The skill (Fig. 13a) of the recalibrated CCLM ensemble has smaller lead-time dependence than the uncalibrated ensemble (cf. Fig. 5). All regions except the British Isles show increased skill. The eastern regions clearly benefit from the calibration (Fig. 12b) especially for lead-time 1–4. This may be related to problems with the regionally initialized soil mainly in Eastern Europe (cf. Section 4.3). The largest improvement is found in Southern Europe (IP, MD). The gain in skill is lower in Central Europe and Scandinavia. For the British Isles the skill is even slightly lower compared to the raw ensembles for longer lead times (Fig. 12b, light blue line) and for some grid points (cf. Fig. 11). For the Iberian Peninsula the area mean added value of recalibration increases over time. For the most other regions, the added value decreases towards the end of the decade.

4.8. Hindcast skill for user-oriented climate indicators and extremes

The above results provided a robust assessment of the skill and added value for near-surface temperature. While the mean temperature is the most commonly used variable for the verification, it is not the climate indicator most useful to potential users. However, there are many temperature-derived indices, where decadal predictions may provide valuable and skilful climate information. This includes hot and cold extremes

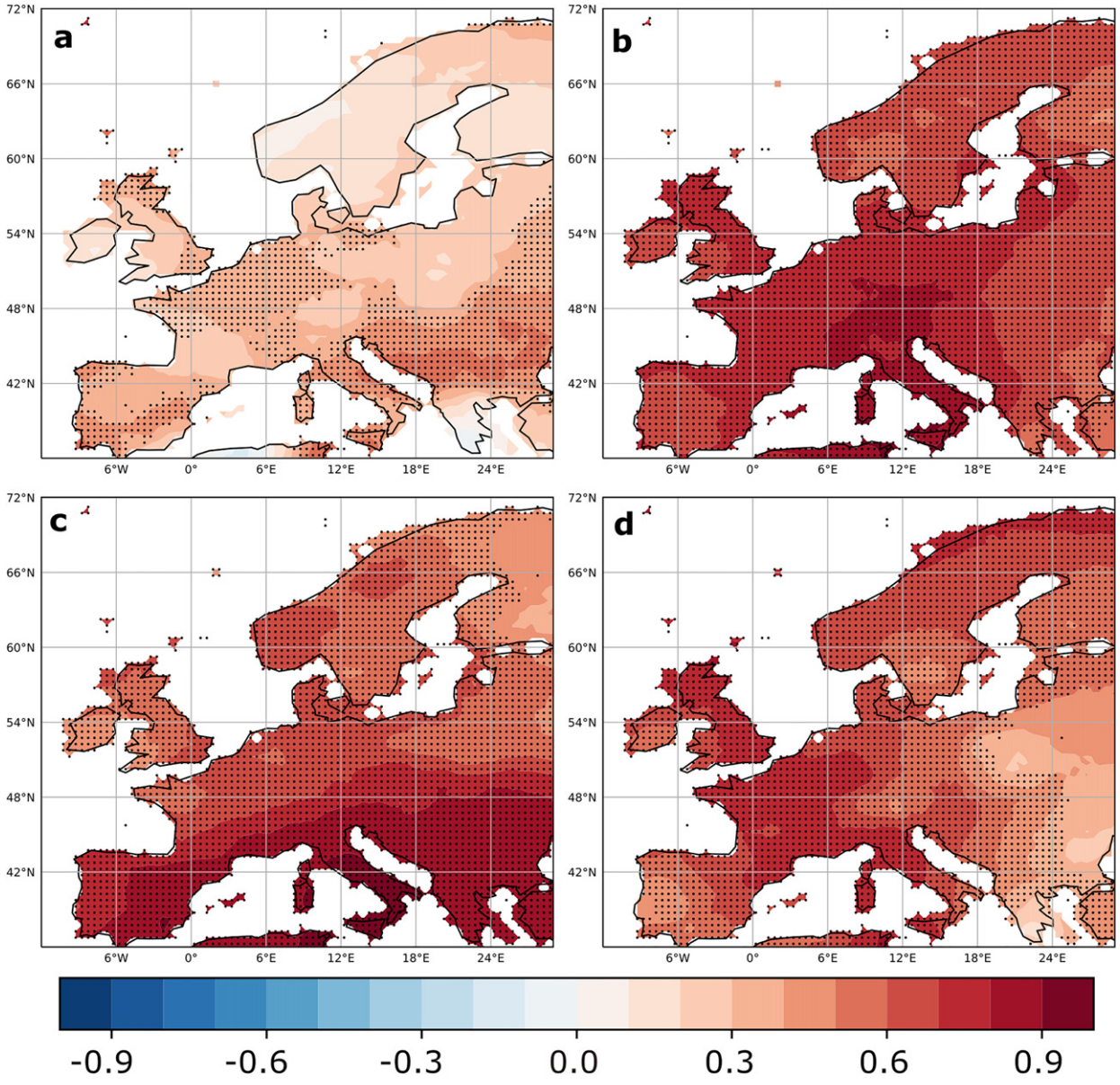

Fig. 9. Correlation of seasonal 4-year mean temperature CCLM b1 lead-years 2–5 for the starting years 1960–2010 (analysis period 1962–2015). Observations CRU TS4.01, for (a) winter (DJF), (b) spring (MAM), (c) summer (JJA) and (d) autumn (SON).

or vegetation based climate indicators (e.g. ETCCDI indices Expert Team of Climate Change Detection Indices, Zhang et al., 2011). Uhlig (2016) found a potential skill for such climate indicators in the regional MiKlip ensemble, especially for indices relying on summer temperatures or indicators using seasonal to annual integrated values. The skill for two types of temperature-derived variables, namely the summer daily maximum temperature and the 'heating degree days (HD)', for the regional hindcasts are displayed in Fig. 13. Both variables are derived from daily data. Therefore, the E-OBS data are used as observational reference. The daily maximum temperatures for JJA are an indicator for summerly heat conditions, which are relevant for heat stress and health issues. There is a high significant MSESS over most of Europe even for the uncalibrated hindcasts (Fig. 13a). The mean skill is comparable to the skill of daily mean temperature in summer if E-OBS is used as reference. The recalibration improves the MSESS significantly over

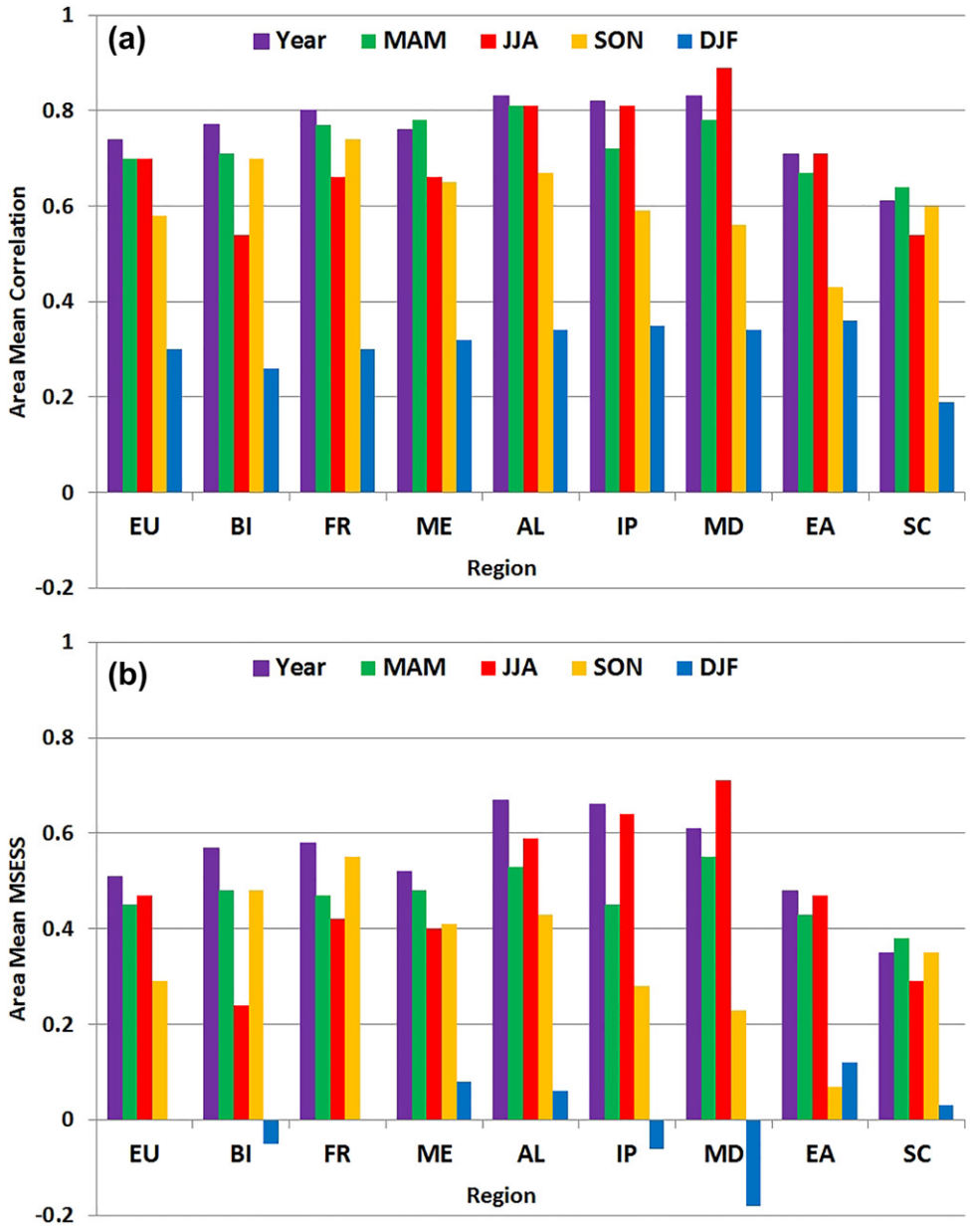

Fig. 10. Area mean skill for 4-year mean temperature CCLM b1 lead-years 2–5 starting years 1960–2010. Correlation (a) and MSESS (b).

most parts of continental Europe, especially in Eastern Europe and Southern Scandinavia. These findings hint at the potential for the prediction of summer extremes over Europe (cf. Eade et al., 2012).

The heating degree-days provide an indicator for the potential demand for domestic heating. It takes those days into account with a mean temperature below a threshold (here 17 °C) and sums up the temperature difference to this threshold as follows:

$$\mathrm{HD} = \sum \max(17\ ^\circ\mathrm{C} - T; 0\ ^\circ\mathrm{C}) \qquad (6)$$

For this variable too, the skill of the uncalibrated hindcasts for lead-years 2–5 is significant for most of Europe and further improved by the recalibration.

These results indicate a skill for different temperature derived climate indicators, which might be suitable for practical applications, in particular after recalibration.

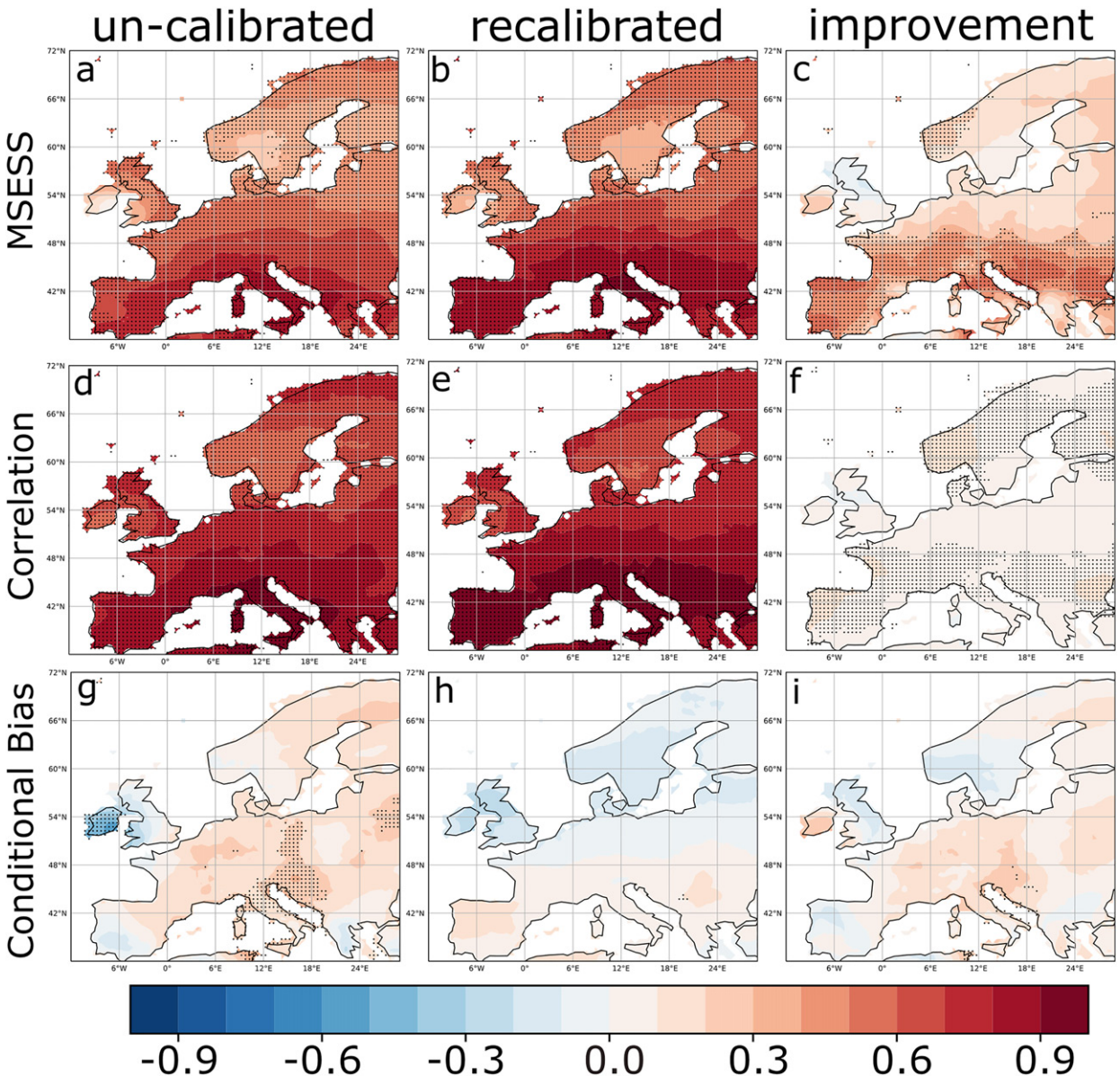

Fig. 11. Comparison skill scores for temperature with and without recalibration. CCLM b1 ensemble, period 1967–2015, lead-time year 2–5. (a, d, g) Un-calibrated data. (b, e, h) Recalibrated. (c, f, i) Added value recalibration (as defined in Section 3). (a–c) MSESS. (d–f) ACC. (g–i) Conditional bias.

5. Summary and conclusions

The objective of this work was to assess and characterize various aspects of regional decadal predictions for Europe based on the MiKlip prediction system. We addressed the spatial and temporal skill patterns on the annual and seasonal time-scale, for different lead- and averaging-times and different hindcast periods. We evaluated the observation uncertainty and the potential for improving the skill by post-processing. The main conclusions are as follows:

- The results presented confirm that there are robustly detectable variations of the predictive skill for Europe at scales finer than the $5° \times 5°$ resolution suggested by Goddard et al. (2013) for temperature. These robust signals indicate that high-resolution decadal predictions for Europe are viable and can

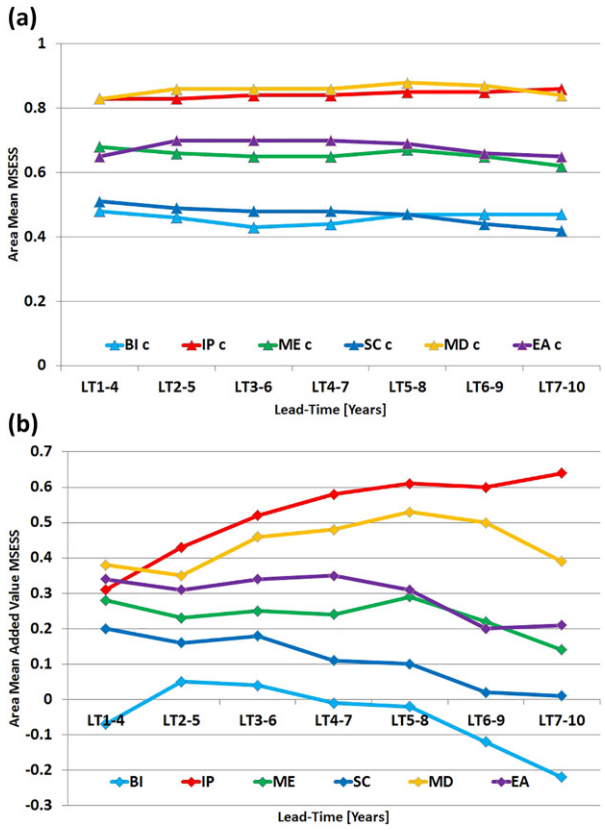

Fig. 12. (a) Area mean MSESS CCLM b1 recalibrated 4-year mean temperature for different PRUDENCE regions and lead-times. Reference: climatology and CRU TS4.01 (cf. Fig. 6 for the un-calibrated hindcasts). (b) Added value MSESS for temperature of the calibrated with the uncalibrated CCLM ensemble as reference.

provide valuable climate information at regional scales.

- The decadal predictions provide a high skill for temperature and a generally low degradation of the scores over lead-time (Section 4.3). This skill arises from the initialization as well as from the climate trend. We showed, that the decadal hindcasts provide a better representation of the decadal variations (Section 4.5) and higher skill scores (Section 4.2) as un-initialized reference simulations.
- The assessment of the added value of downscaling compared to the global forcing data (Section 4.2) confirmed the findings by Mieruch et al. (2014) and Reyers et al. (2017), that the regionalization is able to reduce the bias and improve the reliability of the hindcasts. The correlation of the global hindcasts is already quite high. Therefore, any further improvement is limited. In some mainly northern European regions, the regionalization even slightly reduces the skill.
- This study provides more robust estimates of the regional pattern of the added value of initialization, with the largest improvement over the Mediterranean region, which is in line with findings from Guemas et al. (2015). Further, regions with improvements are Scandinavia and Central to Eastern Europe. In the westernmost part of Europe the gain by the initialized hindcasts is less systematic. Further studies are necessary to understand why the un-initialized simulations already provide such a high skill in these regions.
- The temporal evolution of the predictive skill indicates a low skill for the early hindcast years after 1960 (Section 4.5). This might be due to the limited amount of dense, high-quality observation data for this period, necessary to derive the initial conditions. For later periods, the skill improves rapidly. Another potential explanation could be that the predictive skill for Europe depends on the state of teleconnection pattern like the AMO. To analyse such dependence, a longer hindcast period is needed, which covers at least one or two cycle of the AMO.
- The added value of initialization in lead-year 1 decreases over time. This does not indicate a reduction in skill of the decadal hindcasts. The un-initialized ensemble is also able to capture the climate trend, which provides a larger contribution to the predictability compared to the decadal variability in recent years.
- The averaging time dependence (Section 4.4) reveals an optimum averaging period between four and seven years, where the skill is sufficiently high and an added value of downscaling can be found over a large part of the domain. It can be concluded that the 4-year mean values suggested by Goddard et al. (2013) and often used in the assessment of the skill of decadal predictions pose a reasonable compromise between the need for temporarily highly resolved climate information and sufficient skill levels.
- The annual cycle of the skill (Section 4.6) shows a maximum for summer in Southern Europe, which is in line with the findings from Guemas et al. (2015) and Müller et al. (2012). Moreover, this pattern does not occur in all regions. There is a shift in the seasonal pattern with the highest skill for the British Isles and France in the West in spring and autumn to a mainly summer and partly spring maximum of the skill scores further to the East. Generally, the annual skill scores are higher than the seasonal ones.
- The choice of the observational reference plays a role in the verification of decadal predictions. By using two long-term Europe-wide available regional gridded observation based data sets, we were able to

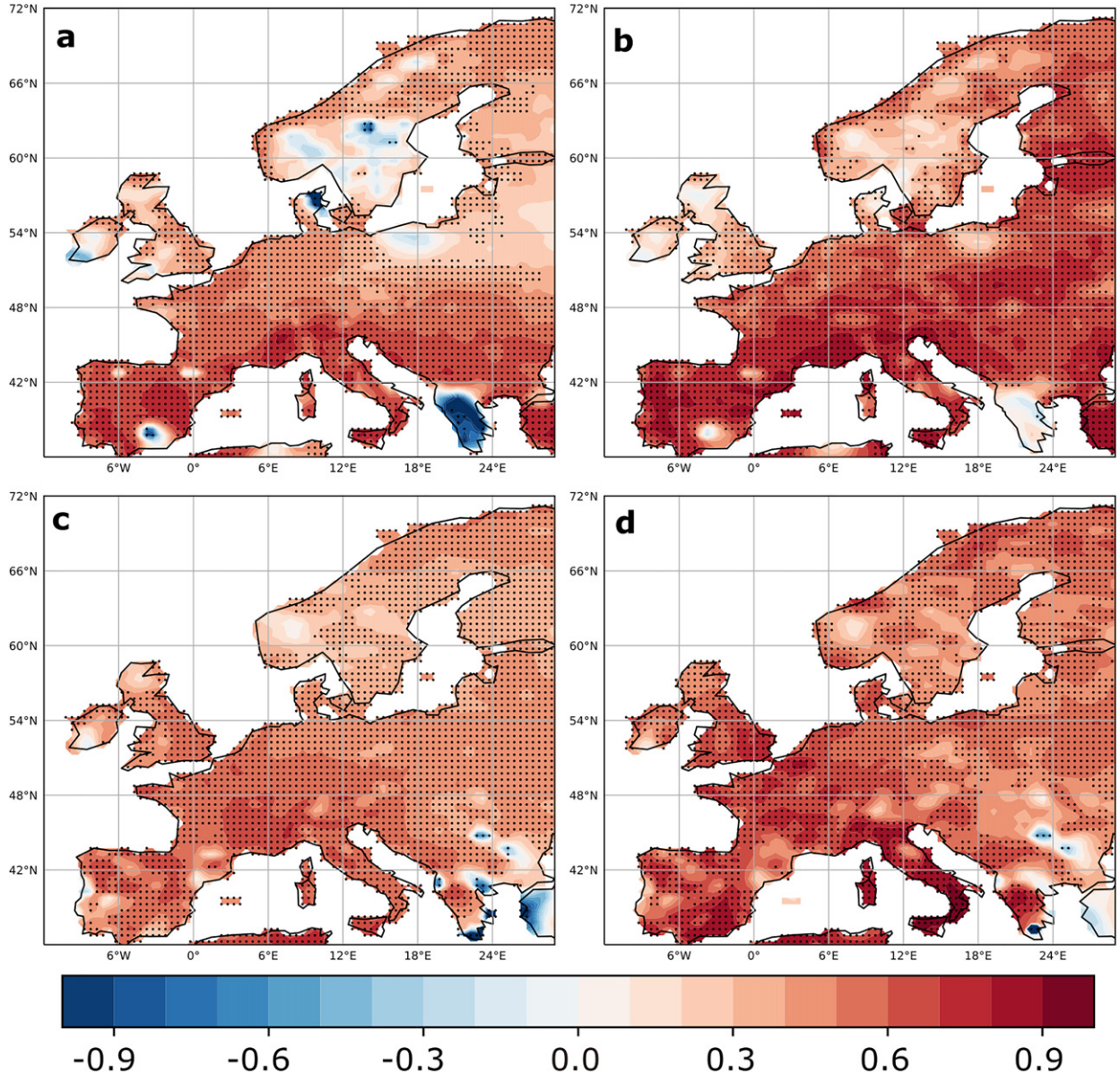

Fig. 13. MSESS for CCLM b1 for the period 1967–2015 lead years 2–5. Daily maximum temperature JJA (a, b) and the heating degree days ($HD = \sum \max(17^{\circ}\mathrm{C} - T; 0^{\circ}\mathrm{C})$, (c, d) un-calibrated (a, c) vs. calibrated data (b, d). Observational reference: E-OBS.

identify regions where a model evaluation is problematic. In general, the verification against both references yield qualitatively similar results (Section 4.1). CRU provides the spatially smoother data set, as it relies on a lower number of stations. The skill scores compared to this reference are often slightly higher. There might be fewer problems with inhomogeneous grid points. Another reason may be related to the 'double penalty' problem in the verification of high-resolution data, where slight displacements of the forecasted position of a signal might strongly affect score like the MSESS (e.g. Mass et al., 2002). This double penalty of higher spatial variability might also affect the estimation of the added value of regional downscaling. For the analysis of extremes, daily data are necessary, which are provided by E-OBS but not by CRU.

- The aspects covered in Section 4.1–4.6 use the uncalibrated output from the hindcasts. It can be shown that a recalibration is able to further improve the

skill (Section 4.7). The applied recalibration method is able to reduce the conditional and unconditional bias and improve the reliability of the decadal predictions. It also reduces the lead-time dependence of the skill scores.

- The MiKlip regional decadal hindcasts provide a high skill for near-surface temperature. The skill also extends to further temperature related variables (Section 4.8). This allows the application of the decadal predictions to user-oriented climate indicators, providing more relevant climate information to the public or the economy for planning or prioritizing actions a few years ahead (cf. Kushnir et al., 2019).

These conclusions give evidence that regional climate predictions can potentially provide valuable climate information on the decadal time-scale. Future work will focus on the assessment of user-oriented climate indicators and the decadal variability of extremes over Europe.

Acknowledgements

The MPI ensemble simulations were provided by the MPI for Meteorology in Hamburg, Germany. All simulations were performed at the German Climate Computing Center (DKRZ). The authors thank the developers of the MiKlip Central Evaluation System – especially Christopher Kadow and Sebastian Illing – for their efforts and support. We acknowledge the E-OBS dataset from the EU-FP6 project ENSEMBLES (http://ensembles-eu.metoffice.com) and the data providers in the ECA&D project (http://www.ecad.eu) and the CRU TS4.01 dataset provided by the Climate Research Unit of the University of East Anglia.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The research programme MiKlip is funded by the German Ministry of Education and Research [BMBF, contract numbers 01LP1518A-E, 01LP1519A, 01LP1520A]. J.G. Pinto thanks the AXA Research Fund for the support. We acknowledge support by Deutsche Forschungsgemeinschaft and Open Access Publishing Fund of Karlsruhe Institute of Technology.

References

Bellucci, A., Haarsma, R., Gualdi, S., Athanasiadis, P. J., Caian, M. and co-authors. 2015. An assessment of a multi-model ensemble of decadal climate predictions. *Clim. Dyn.* 44, 2787–2806. doi:10.1007/s00382-014-2164-y

Ding, R., Li, J., Zheng, F., Feng, J. and Liu, D. 2016. Estimating the limit of decadal-scale climate predictability using observational data. *Clim. Dyn.* 46, 1563–1580. doi:10.1007/s00382-015-2662-6

Doblas-Reyes, F. J., Andreu-Burillo, I., Chikamoto, Y., García-Serrano, J., Guemas, V. and co-authors. 2013. Initialized near-term regional climate change prediction. *Nat. Commun.* 4, 1715. doi:10.1038/ncomms2704

Doms, G. and Schättler, U. 2002. *A description of the non-hydrostatic regional model LM, Part I: Dynamics and Numerics*, Tech. rep. Deutscher Wetterdienst, P.O. Box 100465, 63004 Offenbach, Germany, LM_F90 2.18.

Eade, R., Hamilton, E., Smith, D. M., Graham, R. J. and Scaife, A. A. 2012. Forecasting the number of extreme daily events out to a decade ahead. *J. Geophys. Res.* 117, D21110.

Goddard, L., Kumar, A., Solomon, A., Smith, D., Boer, G. and co-authors. 2013. A verification framework for interannual to- decadal predictions experiments. *Clim. Dyn.* 40, 245–272. doi:10.1007/s00382-012-1481-2

Guemas, V., García-Serrano, J., Mariotti, A., Doblas-Reyes, F. and Caron, L. 2015. Prospects for decadal climate prediction in the Mediterranean region. *Q.J.R. Meteorol. Soc.* 141, 580–597. doi:10.1002/qj.2379

Harris, I., Jones, P. D., Osborn, T. J. and Lister, D. H. 2014. Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset. *Int. J. Climatol.* 34, 632–642.

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D. and co-authors. 2008. A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *J. Geophys. Res. Atmos.* 113, D20119. doi:10.1029/2008JD010201

Hermanson, L., Eade, R., Robinson, N. H., Dunstone, N. J., Andrews, M. B., Knight, J. R., Scaife, A. A. and Smith, D. M. 2014. Forecast cooling of the Atlantic subpolar gyre and associated impacts. *Geophys. Res. Lett.* 41, 5167–5174. doi:10.1002/2014GL060420

Hofstra, N., Halock, M., New, M. and Jones, P. D. 2011. Testing E-OBS European high-resolution gridded data set of daily precipitation and surface temperature. *J. Geophys. Res.* 114, D21101.

ICPO. 2011. Data and Bias Correction for Decadal Climate Predictions. CLIVAR Publication Series No. 150. 6 pp.

Illing, S., Kadow, C., Kunst, O. and Cubasch, U. 2014. MurCSS: a tool for standardized evaluation of decadal hindcast systems. *J. Open Res. Software* 2, e24. DOI:

Jacob, D., Petersen, J., Eggert, B., Alias, A., Christensen, O. B. and co-authors. 2014. EURO-CORDEX: new high-resolution climate change projections for European impact research. *Reg. Environ. Change* 14, 563–578. doi:10.1007/s10113-013-0499-2

Jungclaus, J. H., Fischer, N., Haak, H., Lohmann, K. and Marotzke, J. 2013. Characteristics of the ocean simulations in MPIOM, the ocean component of the MPI-Earth system

model. *J. Adv. Model. Earth Syst.* 5, 422–446. doi:10.1002/jame.20023

Kadow, C., Illing, S., Kunst, O., Rust, H. W., Pohlmann, H. and co-authors. 2016. Evaluation of forecasts by accuracy and spread in the MiKlip decadal climate prediction system. *Metz.* 25, 631–643. doi:10.1127/metz/2015/0639

Keenlyside, N. S., Latif, M., Jungclaus, J., Kornblueh, L. and Roeckner, E. 2008. Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature* 453, 84–88. doi:10.1038/nature06921

Khodayar, S., Sehlinger, A., Feldmann, H. and Kottmeier, C. 2015. Sensitivity of soil moisture initialization for decadal predictions under different regional climatic conditions in Europe. *Int. J. Climatol.* 35, 1899–1915. doi:10.1002/joc.4096

Kim, H. M., Webster, P. J. and Curry, J. A. 2012. Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts. *Geophys. Res. Lett.* 39, L10701.

Kothe, S., Tödter, J. and Ahrens, B. 2016. Strategies for soil initialization of regional decadal climate predictions. *Meteorol. Z* 25, 773–794.

Kushnir, Y., Scaife, A. A., Arritt, R., Balsamo, G., Boer, G. and co-authors. 2019. Towards operational predictions of the near-term climate. Nat. Clim. Change.

Latif, M. and Keenlyside, N. 2011. A perspective on decadal climate variability and predictability. *Deep Sea Res. Part II: Top. Stud. Oceanogr.* 58, 1880–1894. doi:10.1016/j.dsr2.2010.10.066

Marotzke, J., Müller, W. A., Vamborg, F. S. E., Becker, P., Cubasch, U. and co-authors. 2016. MiKlip – a national research project on decadal climate prediction. *Bull. Amer. Meteorol. Soc.* 97, 2379–2394. doi:10.1175/BAMS-D-15-00184.1

Matei, D., Pohlmann, H., Jungclaus, J., Müller, W., Haak, H. and co-authors. 2012. Two tales of initializing decadal climate prediction experiments with the echam5/mpi-om model. *J. Clim.* 25, 8502–8523. doi:10.1175/JCLI-D-11-00633.1

Mass, C. F., Ovens, D., Westrick, K. and Colle, B. A. 2002. Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteorol. Soc.* 83, 407–430. doi:10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G. and co-authors. 2009. Decadal prediction. *Bull. Amer. Meteorol. Soc.* 90, 1467–1485. doi:10.1175/2009BAMS2778.1

Mieruch, S., Feldmann, H., Schädler, G., Lenz, C.-J., Kothe, S. and co-authors. 2014. The regional MiKlip decadal forecast ensemble for Europe: the added value of downscaling. *Geosci. Model Dev.* 7, 2983–2999. doi:10.5194/gmd-7-2983-2014

Moemken, J., Reyers, M., Buldmann, B. and Pinto, J. G. 2016. Decadal predictability of regional scale wind speed and wind energy potentials over Central Europe. *Tellus A: Dyn. Meteorol. Oceanogr.* 68, 29199. doi:10.3402/tellusa.v68.29199

Müller, W. A., Baehr, J., Haak, H., Jungclaus, J. H., Kröger, J. and co-authors. 2012. Forecast skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology. *Geophys. Res. Lett.* 39, L22707.

Müller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M. and co-authors. 2018. A higher-resolution version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR). *J. Adv. Model. Earth Syst.* 10, 1383–1413. doi:10.1029/2017MS001217

Murphy, A. H. 1988. Skill scores based on the mean squared error and their relationships to the correlation coefficient. *Mon. Wea. Rev.* 116, 2417–2424. doi:10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2

Pasternack, A., Bhend, J., Liniger, M. A., Rust, H. W., Müller, W. A. and co-authors. 2018. Parametric decadal climate forecast recalibration (DeFoReSt 1.0). *Geosci. Model Dev.* 11, 351–368. doi:10.5194/gmd-11-351-2018

Pohlmann, H., Jungclaus, J. H., Köhl, A., Stammer, D. and Marotzke, J. 2009. Initializing decadal climate predictions with the GECCO oceanic synthesis: effects on the North Atlantic. *J. Clim.* 22, 3926–3938. doi:10.1175/2009JCLI2535.1

Pohlmann, H., Smith, D. M., Balmaseda, M. A., Keenlyside, N. S., Masina, S. and co-authors. 2013. Predictability of the mid-latitude Atlantic meridional overturning circulation in a multi-model system. *Clim. Dyn.* 41, 775–785. doi:10.1007/s00382-013-1663-6

Reyers, M., Feldmann, H., Mieruch, S., Pinto, J. G., Uhlig, M. and co-authors. 2019. Development and prospects of the regional MiKlip decadal prediction system over Europe: predictive skill, added value of regionalization, and ensemble size dependency. *Earth Syst. Dynam.* 10, 171–187. doi:10.5194/esd-10-171-2019

Rockel, B., Will, A. and Hense, A. 2008. The regional climate model COSMO-CLM (CLM). Editorial. *Meteorol. Z* 12, 347–348.

Romanova, V., Hense, A., Wahl, S., Brune, S. and Baehr, J. 2018. Skill assessment of different ensemble generation schemes for retrospective predictions of surface freshwater fluxes on inter and multi annual timescales. *Meteorol. Z.* 27, 111–124. doi:10.1127/metz/2017/0790

Smith, D. M., Cusack, S., Colman, A. W., Folland, C. K., Harris, G. R. and co-authors. 2007. Improved surface temperature prediction for the coming decade from a global climate model. *Science* 317, 796–799. doi:10.1126/science.1139540

Smith, D. M., Scaife, A. A., Boer, G. J., Caian, M., Doblas-Reyes, F. J. and co-authors. 2013. Real-time multi-model decadal climate predictions. *Clim. Dyn.* 41, 2875. doi:10.1007/s00382-012-1600-0

Stevens, B., Giorgetta, M. A., Esch, M., Mauritsen, T., Crueger, T. and co-authors. 2013. Atmospheric component of the MPI-M earth system model: ECHAM6. *J. Adv. Model. Earth Syst.* 5, 146–172. doi:10.1002/jame.20015

Sutton, R. T. and Hodson, D. L. 2005. Atlantic Ocean forcing of North American and European summer climate. *Science* 309(5731), 115–118. doi:10.1126/science.1109496

Uhlig, M. 2016. *Regional decadal climate predictions for Europe – Feasibility & Skill*. PhD-Thesis, KIT Karlsruhe.

van der Schrier, G., van den Besselaar, E. J. M., Klein Tank, A. M. G. and Verwer, G. 2013. Monitoring European average temperature based on the E-OBS gridded data set.

J. Geophys. Res. Atmos. 118, 5120–5135. doi:10.1002/jgrd.50444

Wilks, D. S. 2011. *Statistical Methods in the Atmospheric Sciences*. San Diego, California: Academic Press, 3rd revised edition.

Zhang, X., Alexander, L., Hegerl, G. C., Jones, P., Tank, A. K. and co-authors. 2011. Indices for monitoring changes in extremes based on daily temperature and precipitation data. *Wires. Clim. Change* 2, 851–870. doi:10.1002/wcc.147