Interactive comment on "Development and prospects of the regional MiKlip decadal prediction system over Europe: Predictive skill, added value of regionalization and ensemble size dependency" by Mark Reyers et al.

Reviewer 1

We thank the Reviewer for his/her thoughtful review and the specific suggestions. As one major point the Reviewer suggested to also show metrics that use the climatology as reference. We agree with the Reviewer that this would enhance the significance of our study. We therefore decided to include a new Figure 3 in the revised manuscript showing the MSESS as in Figure 2 but with the climatology as reference instead of the uninitialised historicals (see also answer to specific comments below). Please also note that we have repeated all calculations without detrending the time series following the suggestions of Reviewer 2. Therefore, not only all Figures have changed in the revised version, but we also had to revise large parts of the main text. Further, as we now use solely time series including the trend we have removed Table 1 and Table 2 in the revised manuscript, which showed the comparison between skill scores as derived from time series with and without trend. Instead, we added a section 4.4 and a new Figure 7 dealing with the effect of the low number of starting dates used in our study on the robustness of our results, again following the suggestions of Reviewer 2. All changes are marked in red in the revised manuscript.
Below point-to-point responses of the authors to all major and specific comments of the Reviewer are given, also in red.

General comments
This is a very good study that focuses on the potential merits of regional downscaling decadal climate predictions over Europe. Specifically, the MiKlip prediction system studied uses the low resolution MPI global decadal hindcast ensemble at T63 resolution and dynamically downscales these hindcasts over Europe using the COSMO-CLM model at 0.22_ horizontal resolution. Two 10 member ensemble regional hindcasts of 5 start dates are examined and verified against observational analyses of surface temperature, precipitation and low level wind using three different skill metrics, MSESS, CRPSS and ACC. The authors examine these metrics to answer the following questions: is there potential for skillful regional predictions in Europe? Does regional downscaling provide added value? and How does the skill of these predictions depend on ensemble size? The first two questions are answered affirmatively and for the last question ensemble size stabilizes the skill metrics MSESS and CRPSS at ten members but ACC skill depends on ensemble size beyond ten members. The manuscript meets all the criteria for publication and needs only minor changes.

Specific comments:
The manuscript could be improved in two ways that would increase the significance of the work. First, although there are significantly large regions in Europe where the skill of the initialized hindcasts is positive, there is also a large region in central Europe where the skill is negative. This is particularly true of the MSESS of temperature. Since the reference is forecast is an uninitialized ensemble of 20th century

simulations this raises the question as to the reason for this negative skill. The answer or some speculation to how it arises should be included in the article. In a similar vein, the authors do not include in their discussion any metrics that use the observed climatological distribution as the reference forecast, so that skill is measured solely using comparison with observations.

Answer: We thank the Reviewer for these helpful comments, which helped to improve our manuscript. Also in the revised Figure 2 the strong negative MSESS over Central Europe is still visible (without detrending the time series, see main comments of Reviewer 2). It is difficult (and also beyond the scope of our study) to find a physical interpretation for this negative skill. However, in order to get a deeper insight to this issue, we have analysed the time series of spatial mean temperature over Prudence 4 (Mid-Europe) for the CCLM hindcasts, the historicals, and the observations (in this case E-OBS). The observed temperature over Mid-Europe strongly increases from dec1960 to dec1970. At the same time CCLM shows a strong decrease, so that it is out of phase of the observations during the first half of the considered period, while the historicals are closer to the observations for this time range. As a consequence, the MSESS using the historicals as reference is strongly negative (as depicted in Fig. 2), although from dec1980 on the temperature curve of CCLM_b1 agrees well to the observations. We decided not to include an extra Figure with respect to this issue, but added a short paragraph in the revised manuscript.
Further, we followed the second suggestion of the Reviewer and performed an additional analysis including the climatology as reference. The new Figure 3 in the revised paper shows the MSESS as in Figure 2, but with the climatology as reference (instead of the uninitialized historicals). The results are in this case "threefold": While for precipitation results look similar in Fig. 2 and Fig, 3, MSESS skill scores for temperature increase and for wind speed mainly decrease when using the climatology instead of the uninitialised historicals. We added a brief discussion on these results in the revised version.

Technical corrections
Pg 2 Yaeger et al should be Yeager et al
A: Has been corrected throughout the manuscript.

Pg 8 stronger scattered should be more strongly scattered
A. We have changed it accordingly.

Interactive comment on "Development and prospects of the regional MiKlip decadal prediction system over Europe: Predictive skill, added value of regionalization and ensemble size dependency" by Mark Reyers et al.

Reviewer 2

Comment of the authors:

We thank the Reviewer for his/her thoughtful review and the specific suggestions. The Reviewer queried two major points: (i) the low number of starting dates used in our study (which has been discussed in detail in the original manuscript) and (ii) the detrending of the time series. As we generally agree with the Reviewer in both points and take the concerns seriously, we decided to include new Figures in the revised manuscript with respect to point (i) (see also answer to main concerns below). Further, we repeated all calculations without detrending the time series as suggested by the Reviewer and replaced the respective Figures in the revised manuscript. As the new calculations impact the majority of our results, a reformulation of most parts of the manuscript is necessary. These changes are marked in red in the revised manuscript. Note, that we have included a new Figure 3 in the revised manuscript. Following the suggestions of Reviewer 1 this Figure shows the MSESS as in Figure 2 but with the climatology as reference instead of the un-initialised historicals. Further, as we now use solely time series including the trend we have removed Table 1 and Table 2 in the revised manuscript, which showed the comparison between skill scores as derived from time series with and without trend. The point-to-point responses of the authors to all major and specific comments of the Reviewer are also given in red in the following.

Anonymous Referee #2

Received and published: 19 December 2017

The paper gives a preliminary assessment of regional decadal prediction skill over Europe based on a high-resolution regional model forced with boundary conditions obtained from the low-resolution, global MiKlip prediction system. I deem the analysis preliminary because the "development and prospects" of the downscaling system are being assessed at a rather early stage when only 5 hindcast start dates have been completed using the regional model. This is a serious shortcoming that calls into question the reliability of skill scores (computed from 5 data pairs) that are used throughout to make statements about the benefits of downscaling for various fields in various European regions.

Two-tier decadal prediction involving regional downscaling is certainly a topic of high interest, but this manuscript has the feel of an internal technical note that documents some preliminary and very mixed results that are still clouded in uncertainty given the limited temporal sampling. Unless it can be shown (perhaps using the MPI baseline systems) that 5 start dates are sufficient to get an accurate estimate for the skill scores and fields of interest, then what is the point of all this? I suspect that 5 start dates is not sufficient, and that the skill scores reported here are very "noisy" as a result. This may contribute to *the mixed results and lack of strong take-away messages from this paper. It may be better to wait until more downscaled start dates have been completed before resubmission of this analysis.*

Answer:

We agree that the consideration of only five starting years lead to “noisier” results. But we argue, that this noisiness affects mainly chapter 4.1, which deals with the skill distribution over Europe. Chapters 4.2 and 4.3, which focus on the added value and the ensemble size dependency respectively, provide robust results even when using two times 40 simulations. Therefore, the major parts of the results are not strongly affected by the sample size issue.

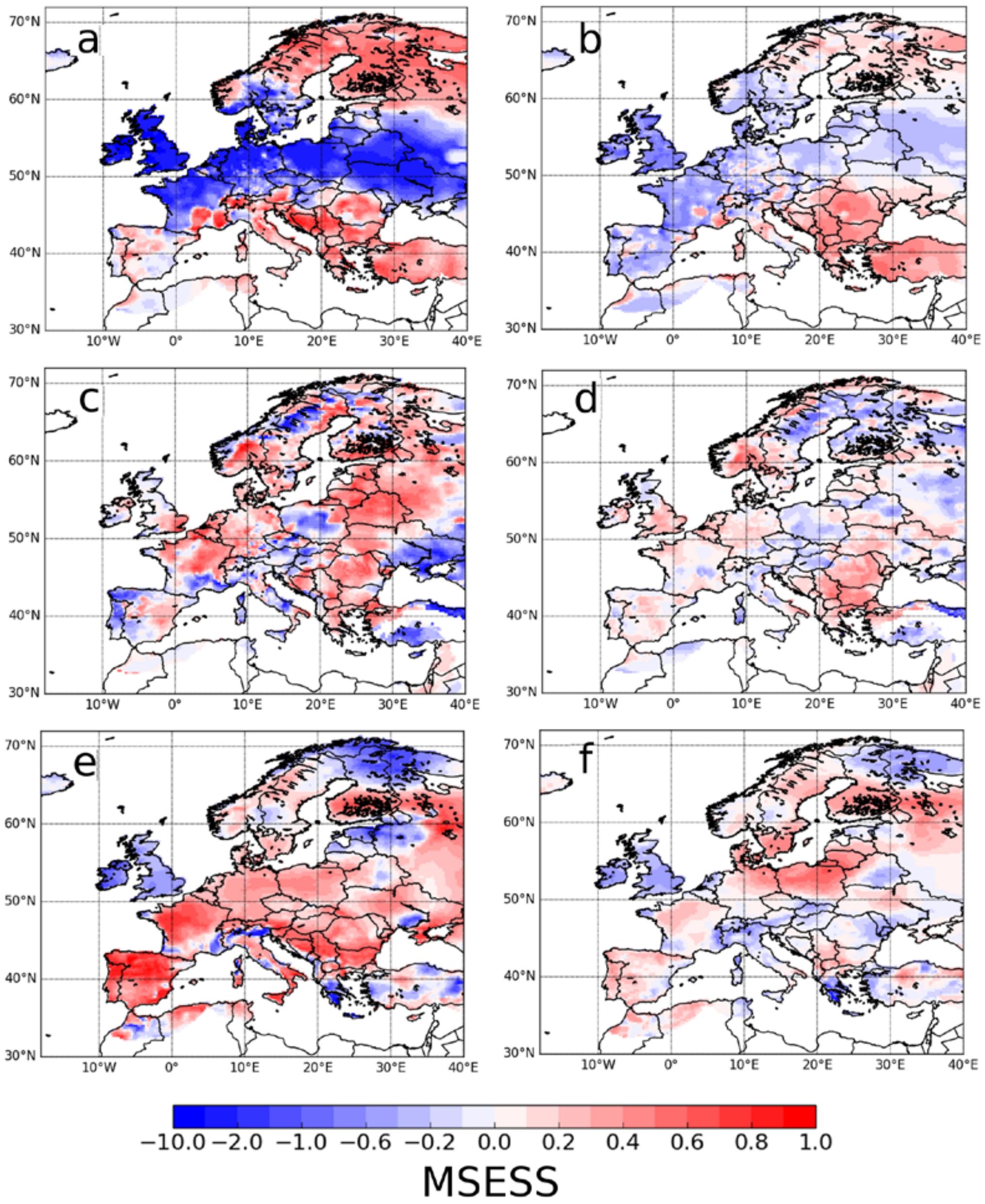


Figure R1: Spatial distribution of the MSESS for the multi-annual mean anomalies of lead years 1-5 in MPI_b1 for (a,b) temperature, (c,d) precipitation, and (e,f) wind speed. For the left panels five start years (dec1960, dec1970, dec1980, dec1990, dec2000) have been used, while for the right panels all start years from dec1960 to dec2000 are taken into account.

Nevertheless, we performed a comparative analysis (as suggested by the reviewer) of the skill estimates for the three variables addressed in the paper derived from a) starting years every 10 years (1960, 1970,..,2000) as in the original manuscript and b) annual starting dates (1960-2000) for the global 10 member ensemble with MPI-ESM-LR baseline1. Baseline1 is the only ensemble used in the paper which provides 10 members throughout the whole hindcast period. The results (see attached Figure R1 showing the MSESS and new Figure 7 in the revised manuscript for the correlation) show a general qualitative agreement, though of course not a quantitative one. As expected, a larger sample size provides smoother skill estimates, less noisy than with the smaller sample size. But in general the findings for most regions that showed hindcast skill in the original manuscript are still correct for the extended ensemble. We have included this additional analysis in the new section 4.4 in the revised paper to point out, how and where the smaller sample size affects the findings and that way putting them into perspective. As we show the MSESS already in Fig. 2 and Fig. 3 of the revised paper we decided to only show the correlation in Fig. 7 but to also discuss the results for the MSESS (see Fig. R1) in the main text.

Another main concern is the use of detrending, which probably exacerbates the sampling issues (how well-defined is a trend computed from 5 data points?). There is no real need to detrend since you have an uninitialized ensemble that allows you to determine the skill improvement relative to the externally-forced signal (yes, pure ACC will be higher, but you can show delta(ACC), i.e. the change in ACC relative to the uninitialized ensemble).

The quality of the writing is decent, but not high, and there are numerous instances of poor English construction (some noted below). A thorough proofreading is in order if this is to be resubmitted.

A: We thank the Reviewer for this helpful comment. We agree that detrending is not necessary when using an uninitialized ensemble as reference. We therefore decided to redo all calculations of our study without detrending of the time series and included the new results in the revised manuscript. The new results generally differ only slightly from the outcomes of the analysis with detrending (cf. Figures in the original version to Figures in the revised manuscript). Altogether, we found a slight improvement. This is the case for both the absolute skill scores of the regional hindcasts (e.g. Fig. 2) and the added value of downscaling (e.g. Fig. 5 and 6). Noticeable differences are e.g. found for wind in Eastern Europe (see Fig. 2e,f). We have changed the text according to the new data processing procedure and to the new figures throughout the text. Additionally, following the Reviewers suggestion, we have included delta(ACC) as a measure for the change in ACC of the hindcasts relative to the uninitialized ensemble in Fig. 5 and 6. Further, we have carefully proofread the revised version of the manuscript as suggested by the Reviewer.

Specific Comments and Questions:

P2,L8: Here and throughout: “Yaeger” should be “Yeager”.

A: Citation changed throughout the revised version.

P2,L11: It’s not clear what the point is of the “while few” construction. Are you contrasting the large number of studies focusing on global metrics with the relatively few studies focusing on storm tracks, etc? Please rewrite.

A: We agree with the Reviewer that this sentence is misleading. We have rephrased it in the revised version.

P2,L13: What is this an example of? Why cite Sutton and Hodson (2005) in a paragraph focused on initialized decadal prediction?

A: We have removed this sentence in the revised manuscript.

P3,L7-18: The motivation for the present work needs to be clarified, particularly since it is not at all clear how the present study differs from the closely related recent MiKlip studies that have just been cited (Kadow et al. 2016; Mieruch et al. 2014; Haas et al. 2016; Moemken et al. 2016).

A: As stated in line 4-6 on page 3, the closely related MiKlip studies are difficult to compare, as they use different skill metrics, pre-processing methods and downscaling approaches. The unique feature of our study is that we use the same methods/metrics not only for the regional but also for the global prediction system. This enables us to give a more general assessment of the prospects of the MiKlip system with respect to basic near surface variables, which affect human life most (temperature, precipitation, and wind speed). However, we agree that this sentence is not sufficient to emphasize our motivation. We have rephrased it in the revised version.

P3,L11: This question is poorly phrased. Do you mean "depend on" the trend or "derive from" the trend?

A: We agree that this question is misleading. We have removed the second part of the question as we only consider time series with trend in the revised version, following the Reviewers suggestion.

P3,L15: This is a repetitive rephrasing of the questions just covered.

A: This paragraph has been removed in the revised manuscript.

P3,L29: I don't understand how ocean temperature and salinity can be nudged towards NCEP/NOAA reanalysis, since the latter is an atmospheric reanalysis.

A: The reviewer is correct. In baseline0 the ocean salinity and temperature anomalies were derived from a simulation with the ocean model MPI-OM forced with the NCEP re-analysis. We have changed the sentence accordingly.

P4,L11: Not clear what is meant by "Analog to the global data"?

A: To avoid misinterpretation, we have changed the text:

"The experiment includes downscaled hindcasts for dec1960, dec1970, dec1980, dec1990, and dec2000, with ten members per decade (hereafter CCLM_b0 and CCLM_b1). The regional ensembles therefore consist of the same time series like the global ensembles MPI_b0 and MPI_b1."

P4,L14: Replace "are" with "is".

A: According to the next Reviewers comment (P4, L16-19) we have changed the paragraph.

P4,L16-19: You already introduced the ERA-driven CCLM simulation in the first line of this paragraph, so consolidate your sentences into one brief description.

A: We have consolidated the sentences in the revised version.

P4,L21: I don't understand what you mean by "uninitialized model simulations started from historical CMIP5 runs". Do you mean downscaled simulations that can be considered "uninitialized" counterparts to CCLM_b0 and CCLM_b1? Do you mean "preindustrial CMIP5 runs"?

A: We agree that this sentence is misleading. We have changed it in the revised version: "To address this issue, uninitialised historical CMIP5 runs are usually considered …".

P4,L32: Replace "the natural variability" with "natural variability". Why use linear detrending to isolate natural variability when you have just introduced an uninitialised ensemble that can be used to quantify the skill associated with external forcing?

A: The forecast skill of a decadal prediction system may origin from two different "processes": a realistic prediction of the long-term trend and a suitable forecast of peaks on inter-annual timescales due to natural variability. To isolate the forecast skill for anomalies on inter-annual time scales, we originally detrended **all** datasets used in this study (as stated in the first paragraph of section 3.1), i.e. not only the hindcasts and the observations, but also the uninitialized historical runs. However, we agree with the Reviewer that detrending is not necessary when we use an uninitialized ensemble that allows us to determine the skill improvement relative to the externally-forced signal (see Reviewer's major comments). Hence, we decided to redo all the analysis without detrending in the revised version (see also our answer to main comments). As a consequence we have removed this paragraph in the revised manuscript.

P5,L14: I think you mean "post-processed time series".

A: No, we mean pre-processed here, as they are processed before they are analysed by using different skill-metrics.

P5,L24: What is the basis for claiming that "skill should originate mainly from the initialization" as opposed to the external forcing? This has not been shown and shouldn't be assumed.

A: We agree with the Reviewer that this hypothesis is too speculative without analysing it in detail. We therefore have removed this clause in the revised version.

P5,L25-: What are F(y) and Fo(y)? Please explain the CRPS equation. What exactly is CDF and how is it computed?

A: The CRPS is defined as the quadratic measure of the discrepancy between the forecast cumulative density function (*F)* and the observed cumulative density function (*Fo)* of a variable *y*. The cumulative density function (CDF) of a real-valued variable *y* is defined as:

$CDF(y) = P(y \leq t)$,

where *P* is the probability that the variable *y* has a value of less than or equal to *t*.

We added this information to the revised manuscript.

P6,L23-P7,L2: This is repetitive.

A: This paragraph has been removed in the revised manuscript.

Fig 2: Suggest using a nonlinear scale for MSESS, such as -9 to 0.9 as in Shaffrey et al. (2016, doi:10.1007/s00382-016-3075-x), because this metric is not symmetric

about 0 in terms of relative improvements in MSE. Please clarify that these are for annual mean (ie, not seasonal mean) anomalies.

A: Following the Reviewers suggestion, we have used a nonlinear scale for Fig. 2 and the new Fig.3 in the revised version and clarified that skill scores are for annual mean anomalies in the Figure captions.

P7,L7: I presume the detrending has been performed similarly for observations and for the uninitialized historical runs? This isn't explicitly mentioned.

A: In the first paragraph of section 3.1 of the original manuscript we stated that all datasets "are pre-processed in an analogous manner". However, as we decided to keep the trend in the datasets for our analysis, we have removed this paragraph in the revised manuscript.

Table 1: What is the meaning of "The uninitialized historical ensemble has been used as reference dataset", given that this is a table of ACC scores? Am I correct that this table displays correlations computed from 5 data points (corresponding to the 5 start years)? Clearly the externally-forced trend is important and so this table should include ACC scores for the uninitialized historical runs for comparison.

A: We thank the Reviewer for this note, which is correct. However, as we keep the trend in the analysed datasets in the revised manuscript as suggested by the Reviewer, we decided to remove Table 1 and Table 2 in the updated version as they originally showed "trend versus detrended" results. With respect to the correlation, we now included ACC scores for the uninitialized historical runs in Figures showing correlation scores.

P7,L11: What is the meaning of "increases" ˇTrelative to uninitialized or relative to detrended?

A: The MSESS for the datasets with trend increases relative to the detrended time series. However, this paragraph has been removed in the revised version (see answer to comment above).

P7,L34: I would say Figure 2 shows more than a "slight shift".

A: This is indeed a too strong generalization of the results. Discrepancies between the two hindcast generations are rather small for temperature, but can be quite large for precipitation and wind speed, depending on the region. We have clarified this in the revised version.

P8,L2: Here and elsewhere delete "exemplary" as it is not being used properly.

A: Following the Reviewer's suggestion we have deleted exemplary throughout the manuscript.

P8,L4: It's curious that Fig 2e agrees so well with Fig 3a, but Fig 2f is so different from Fig 3b. Can you offer any explanation? In my mind, it calls into question the significance of skill scores computed from 5 data points.

A: It is difficult to find an explanation for this issue. Aside from statistical reasons this might also be related to the detrending of the data. However, in the revised version we included a new Fig. 3 showing the MSESS with the climatology as reference.

P8,L25-29: This discussion begs the question of why you are doing any detrending at all (see comment above)? The purpose of the uninitialized ensemble is precisely to allow you to discriminate between greenhouse-gas induced variability (including trends)

and natural variability (including AMO-related trends). Detrending is confusing matter sâˇAˇ Tjust compared initialized to uninitialized skill.
A: Again, we agree with the Reviewer in this point and have repeated all calculations without detrending. Therefore, this paragraph has been removed in the revised manuscript.

P9,L8: Change “whereas” to “and”.
A: Has been changed.

P9,L9-11: This incomprehensible sentence needs a rewrite.
A: We have rephrased this sentence in the revised version.

P11,L5: I don’t understand this sentence.
A: If we would de-bias the CRPSS this would imply a different processing of the analysed datasets compared to the MSESS and the ACC and would make it difficult to compare the skill analysis. As stated in the introduction and in the response to the 4th specific comment of the Reviewer, this is exactly what we intend to avoid in our study. We therefore decided not to use a de-biased version of the CRPSS. However, as this is obviously stated incomprehensible, we have rephrased this sentence in the revised version.

P11,L24-28: This is because you are doing bootstrapping without replacement; if you allow replacement, then the spread does not necessarily diminish with ensemble size.
A: As already stated in the original manuscript our aim was not to do a bootstrapping, but to do permutations over all useful ensemble combinations. In our opinion the individual n-member ensembles should contain each ensemble member only once. Otherwise, the 10 member ensemble may in an extreme case consist of 10 times the same member, which in our opinion makes no sense for decadal prediction purposes. We would therefore keep the method without replacement in the revised version depending on the Editors decision. However, it is correct that the permutation without replacement results in a decline of the spread with increasing number of members. We have added this information to the paragraph in the revised manuscript.

Development and prospects of the regional MiKlip decadal prediction system over Europe: Predictive skill, added value of regionalization and ensemble size dependency

Mark Reyers[1], Hendrik Feldmann[2], Sebastian Mieruch[2,3], Joaquim G. Pinto[2], Marianne Uhlig[2,4], Bodo Ahrens[5], Barbara Früh[6], Kameswarrao Modali[7], Natalie Laube[2], Julia Mömken[1,2], Wolfgang Müller[7], Gerd Schädler[2], Christoph Kottmeier[2]

[1]Institute for Geophysics and Meteorology, University of Cologne, Cologne, Germany
[2]Institute for Meteorology and Climate Research (IMK-TRO), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[3]Alfred-Wegener Institute for Polar and Marine Sciences, Bremerhaven, Germany
[4]School of Geography, Environment and Earth Sciences, Victoria University of Wellington, Wellington, New Zealand
[5]Institute for Atmospheric and Environmental Sciences, Goethe-University Frankfurt a.M., Frankfurt a.M., Germany
[6]Deutscher Wetterdienst (DWD), Offenbach, Germany
[7]Max Planck Institute for Meteorology, Hamburg, Germany

Correspondence to: M. Reyers, (mreyers@meteo.uni-koeln.de)

Abstract. The current state of development and prospects of the regional MiKlip decadal prediction system for Europe are analysed. The Miklip regional system consists of two 10-member hindcast ensembles computed with the global coupled model MPI-ESM-LR downscaled for the European region with COSMO-CLM to a horizontal resolution of 0.22° (~25km). Prediction skills are computed for temperature, precipitation, and wind speed using E-OBS and an ERA-Interim driven COSMO-CLM simulation as verification datasets. Focus is given to the eight European PRUDENCE regions and to lead years 1-5 after initialization. Evidence of the general potential for regional decadal predictability for all three variables is provided. For example, the initialized hindcasts outperform the uninitialized historical runs for some key regions in Europe, particularly in Southern Europe, and for some variables both in terms of accuracy and reliability. However, forecast skill is not detected in all cases, but it depends on the variable, the region, and the hindcast generation. A comparison of the downscaled hindcasts with the global MPI-ESM-LR runs reveals that the MiKlip prediction system may distinctly benefit from regionalization, in particular for parts of Southern Europe and for Scandinavia. The forecast accuracy and the reliability of the MiKlip ensemble is systematically enhanced when the ensemble size is stepwise increased, and a number of 10 members is found to be suitable for decadal predictions. This result is valid for all variables and European regions in both the global and regional MiKlip ensemble. The present results are encouraging towards the development of a regional decadal prediction system.

1. Introduction

In recent years, the interest in climate predictions on time-scales from one year up to a decade has increased in the climate science community, since this time span falls within the planning horizon for a wide variety of decision makers (Meehl et al., 2009; 2014). A large ensemble of initialised decadal hindcasts has been consolidated in a component of the Coupled Model Intercomparison Project Phase 5 (CMIP5; Taylor et al., 2012), and the number of studies aiming at decadal predictions has strongly increased in recent years (for a review see Meehl et al., 2014). Typically, the North Atlantic is a key region for decadal predictions and forecast skill is found for various quantities such as heat content and SST (e.g. Kröger et al, 2012; Yeager et al., 2012), CO2 uptake (Li et al., 2016) and integrated quantities such as the AMOC (Pohlmann et al., 2013a) and the sub-polar gyre (Matei et al., 2012; Yeager et al., 2012; Robson et al., 2013). Other studies focus on primary meteorological parameters on the global scale, in particular surface temperature (e.g Chikamoto et al., 2012; Doblas-Reyes et al., 2013; Ho et al., 2013; Corti et al., 2015). Comparatively few studies analyse storm tracks (Kruschke et al., 2014, 2016), Atlantic tropical cyclones (Dunestone et al., 2011), intense or extreme events (e.g. Benestad and Mezghani, 2015) or zoom into a certain region of the world (e.g. Guemas et al., 2015).

In the German research consortium MiKlip (http://www.fona-miklip.de), a global decadal prediction system was developed based on the Max-Planck-Institute Earth System Model (MPI-ESM) (for an overview see Marotzke et al., 2016). Within the the project, several hindcast generations were produced. The first two are discussed in this paper. The skill of the MiKlip System for decadal predictions was analysed in a wide variety of recent studies. For example, Müller et al. (2012) investigated global surface air temperature in the first generation of the global MiKlip system (baseline0, which was a contribution to CMIP5) and found that the initialized hindcasts have predictive skill over the North-Atlantic region, while negative skill scores are identified for the tropics. A modified initialization in the second global MiKlip system generation (baseline1) considerably improves the performance in the tropics, but brings only limited skill improvement over the North Atlantic and Europe (Pohlmann et al., 2013b). Kruschke et al. (2014) identified significant positive skill scores for cyclone frequencies over the central North Atlantic in the global baseline0 and baseline1 generations, while no significant skill was detected over the eastern North Atlantic and Europe. Furthermore, Kadow et al. (2016) evaluated the global MiKlip system with respect to temperature and precipitation, giving evidence that an enlargement of the hindcast ensemble generally leads to an improvement of the prediction system.

The MiKlip consortium is to our best knowledge the first institution worldwide which has established a decadal prediction system for the regional scale. With this aim, considerable efforts were made to downscale the global MPI-ESM hindcasts by developing and/or employing different regionalisation techniques. Previous experiences reveal that a skill for regional decadal predictions exists but that the interpretation of the results is quite complex due to the non-linear relationship to the global prediction skill. For example, Mieruch et al. (2014) found rather heterogeneous predictive skill for precipitation and temperature over Europe in the baseline0 generation. The skill differs over space, season, variable, and lead time after initialisation. However, a general feature is an improved model spread for precipitation in the downscaled hindcasts when compared to their global counterparts. A potential for predicting regional peak winds and wind energy potentials over Central Europe several years ahead was identified in Haas et al. (2016) and Moemken et al. (2016). Particularly, they found highest skill scores for the first years after initialisation. All the individual studies analysing the MiKlip prediction system consider different ensembles, variables, lead times, skill metrics, and/or downscaling and data pre-processing methods. Therefore it is difficult to draw general conclusions for the decadal predictability over Europe in the MiKlip decadal prediction system. In particular, an overall statement for the benefit of regionalisation and thus for the prospects of a regional decadal prediction system is hardly possible so far. This motivated us to analyse both the global and the downscaled MiKlip ensemble with respect to different issues.

In this study, the decadal predictive skill for temperature, precipitation, and wind speed over Europe is analysed for the baseline0 and baseline1 generation of the MiKlip system. With this aim, we used the same methodologies for all three variables to ensure comparability. Global MPI-ESM and downscaled hindcast ensembles are considered to address the following four key questions:

- Is there a potential for skilful regional decadal predictions in Europe?
- Does regional downscaling provide an added value for decadal predictions?
- Does the regional decadal predictive skill depend on the ensemble size?
- How does the sample size affect the skill estimates?

The datasets used in this study are described in section 2, followed by the methodologies for data pre-processing and skill analysis in section 3. The results for the four key questions are shown in section 4. A summary and discussion, as well as an outlook for future work are given in section 5.

2. Data

The analysed global hindcasts were simulated with the coupled model MPI-ESM in low-resolution (MPI-ESM-LR; Giorgetta et al., 2013). Its atmospheric component is based on the ECHAM6 model (Stevens et al., 2013) with a horizontal resolution of T63 and 47 vertical levels, which is coupled to the MPI-OM ocean model (Jungclaus et al., 2013) with a horizontal resolution of 1.5° and 40 vertical levels. Two hindcast generations are considered here, both computed with the MPI-ESM-LR but with different initialisation strategies. The first generation (baseline0; Müller et al., 2012) is initialised with oceanic conditions from a coupled experiment, where ocean temperature and salinity anomalies from the NCEP/NOAA reanalysis (Kalnay et al., 1996) were assimilated into the ocean model MPI-OM. For the second generation (baseline1; Pohlmann et al., 2013b), temperature and salinity anomalies from the ocean reanalysis system 4 (ORAS4; Balmaseda et al., 2013) are used instead, together with a full-field 3-D atmospheric initialisation using fields from ERA40 (Uppala et al., 2005) and ERA-Interim (Dee et al., 2011). For both generations, yearly initialised hindcasts are available, each of them comprising a 10-year period. For each starting date, an ensemble was generated using a 1-day lagged initialisation from the assimilation experiments (cf. Marotzke et al., 2016 for more details). For baseline0 there are 10 members for each fifth year and three members for the other years, whereas baseline1 provides 10 members for each starting year. The downscaling experiment was performed with the global forcing from hindcasts of five starting dates are used (1 January 1961, 1971, 1981, 1991, and 2001; hereafter referred to as dec1960, dec1970, dec1980, dec1990, and dec2000) to cover the whole period from 1961-2010. This resulted in an ensemble of 50 global hindcasts per generation (baseline0 and baseline1; hereafter MPI_b0 and MPI_b1).

The global hindcasts are dynamically downscaled to the EURO-CORDEX domain (Giorgi et al., 2006; cf Figure 1) at a horizontal grid resolution of 0.22° using the mesoscale non-hydrostatic regional climate model COSMO-CLM (CCLM; Rockel et al., 2008) on a rotated grid. The model version COSMO4.8-clm17 is employed. By using the MPI-ESM-LR ensemble as driving data, the global "initial condition" perturbation strategy is simply passed to the regional model. The downscaling experiment includes hindcasts for dec1960, dec1970, dec1980, dec1990, and dec2000, with ten members per decade (hereafter CCLM_b0 and CCLM_b1). The regional ensembles therefore consist of the same time series like the global ensembles MPI_b0 and MPI_b1.

We evaluate the performance of both the global MPI-ESM and the regional CCLM hindcasts with the following datasets: For temperature and precipitation we consider the observational dataset E-OBS (Haylock et al., 2008) based on the ECA&D (European Climate Assessment & Dataset; http://eca.knml.nl/) at a regular 0.25°x0.25° grid. As no gridded dataset is available for wind, a CCLM simulation forced with boundary conditions from ERA40 and ERA-Interim is employed as verification dataset for wind speed. For this reanalysis driven simulation, CCLM is applied in the same model setup as for the regionalisation of the global hindcast ensemble (see above).

In this study, we want to quantify if the initialisation with observed climate states improves the performance of decadal predictions. To address this issue, uninitialised historical CMIP5 runs are usually considered as reference dataset (see also section 3.2). With this aim, a 10-member ensemble of uninitialised MPI-ESM-LR historical runs started from a pre-industrial control simulation are used, which are only forced by the aerosol and greenhouse gas concentrations for the period 1850-2005 (e.g. Müller et al., 2012).

3. Methods

3.1 Data processing

All datasets considered in this study are pre-processed in an analogous manner to enable a direct comparison. First, all data are interpolated to the same regular 0.25°x0.25° grid, which corresponds to the resolution of the E-OBS data. At each grid point, monthly anomaly time series are computed by subtracting the long-term means for the period 1961-2010 from the interpolated raw datasets. Finally, annual values are derived and multi-annual means for lead years 1-5 are built for further evaluation.

Following the suggestion of Goddard et al. (2013), the skill analysis is mainly performed for spatial means. Spatial averaging of the anomaly time series is performed for eight PRUDENCE regions over Europe (see Fig. 1; Christensen and Christensen, 2007). Note that we only used grid points over land surfaces for the spatial means, as E-OBS data are not available over the oceans. Additionally, we calculated the predictive skill on the basis of all individual grid points for specific analysis.

3.2 Skill metrics

The following three metrics are used to evaluate the performance of the global and regional hindcast ensembles and to address the four key questions: the continuous ranked probability skill score (CRPSS), the mean squared error skill score (MSESS), and the anomaly correlation coefficient (ACC). The skill metrics are applied to the pre-processed time series described in section 3.1 and are computed for multi-annual means for lead time years 1-5 after initialisation. Recent studies analysing the MiKlip decadal prediction system demonstrated that the MiKlip ensemble performs best for the first years after initialisation for a wide range of variables, while the skill diminishes for longer forecast periods. For example, Müller et al. (2012) found highest skill scores for years 1-4 and 2-5 for annual mean surface temperature both for the North Atlantic region and global means. The same is true for annual wind speed and wind energy potentials over Central Europe, for which skilful predictions are mainly restricted to the first years after initialisation (years 1-4), while negative skill scores are found for longer lead time periods (Moemken et al., 2016). Kruschke et al. (2014) provided evidence that the prediction skill for winter cyclones over the North Atlantic region is best for years 2-5 and reduced for longer time periods. Following the recommendation by Goddard et al. (2013), we focus in the following on the lead time years 1-5 after initialisation.

The CRPSS (e.g. Goddard et al., 2013) is often used to assess the reliability of probabilistic forecast models and defined as

$$CRPSS = 1 - \frac{CRPS_{hind}}{CRPS_{ref}}$$

with

$$CRPS = \int_{-\infty}^{\infty} [F(y) - F_o(y)]^2 \, dy$$

The CRPS is the continuous ranked probability score (Wilks, 2011) and is defined as the quadratic measure of the discrepancy between the forecast cumulative density function (*F*) and the observed cumulative density function (F_o) of a variable y. The cumulative density function (CDF) of a real-valued variable y is defined as:

CDF(y) = P(y ≤ t),

where P is the probability that the variable *y* has a value of less than or equal to *t*. $\mathrm{CRPS}_{\mathrm{hind}}$ is the CRPS of the initialised hindcasts, and $\mathrm{CRPS}_{\mathrm{ref}}$ is the CRPS of a reference dataset, which are in this study the uninitialized MPI-ESM-LR historical simulations. In case of a positive CRPSS the reliability in terms of the probabilistic quality of the forecast spread is higher in the initialised hindcasts than in the reference dataset. It can thus be used to test if the model ensemble spread adequately represents the forecast uncertainty.

The deterministic MSESS (Goddard, 2013) is defined as

$$MSESS = 1 - \frac{MSE_{hind}}{MSE_{ref}}$$

with

$$MSE = \frac{1}{N}\sum_{n}^{N}(\overline{X}_\iota - \ O_i)^2$$

where MSE$_{\text{hind}}$ is the mean squared error (MSE) between the ensemble mean of the initialised hindcasts (X_i) and the verification data, and MSE$_{\text{ref}}$ is the mean squared error of a reference dataset (here: uninitialised historical simulations) versus the verification data (O_i). A positive MSESS means that the hindcasts are closer to the verification dataset than the uninitialised runs, indicating that the initialisation leads to higher accuracy in predicting observed values. In a sensitivity study we additionally choose the climatology as reference dataset. Note that independently from the ensemble size of the hindcast ensembles, the same historical 10-member ensemble is always used as reference dataset for the computation of CRPSS and MSESS.

The ACC (e.g. Wilks, 2011) is computed as the Pearson correlation between the ensemble mean of the hindcasts at a certain location i and the corresponding observations (Obs):

$$ACC_i = \frac{1}{N}\frac{\sum_t hind_t\ Obs_t}{\sigma_{hind}\ \sigma_{Obs}}$$

where $t = 1, \ldots, N$ is the time index. The ACC quantifies the accuracy of the predictions only in terms of the temporal course, while it is independent from the mean bias. To compare the performance of the hindcasts and of the uninitialized historical runs, we compute the difference of the ACC of the hindcasts minus the ACC of the historicals for several issues (hereafter delta_ACC).

4. Results

4.1 Is there a potential for skilful regional decadal predictions in Europe?

In this section we address the first key question and analyze the general potential for skilful regional decadal predictions over Europe. Fig. 2 shows MSESS plots for temperature, precipitation and surface wind speed in CCLM_b0 and CCLM_b1 with the un-initialized simulations as reference. For temperature (Fig. 2a and 2b), positive skill scores are found in both ensembles over Scandinavia and for South-eastern Europe, while a stripe of negative values occurs over the British Isles and Central Europe. The analysis of the time series for Mid-Europe (spatial mean over Prudence region 4) reveals that this negative skill mainly results from a strong temperature increase from dec1960 to dec1970 in the observations, while CCLM_b0 and CCLM_b1 depict a decrease in temperature (not shown), which in fact was observed in Southern Europe for instance. As a consequence, the temperature in the hindcasts has a larger bias than the uni-initialized simulations compared to the observations during the first half of the considered period, but agree well to the observations from dec1980 onwards. The largest deviations between CCLM_b0 and CCLM_b1 are found for Iberia, parts of southern France and Italy, where the MSESS is positive for CCLM_b1 but neutral to negative for CCLM_b0.

Deviations between both ensembles are larger for precipitation (Fig. 2c and 2d), where the MSESS fields are distinctly patchier when compared to temperature (Fig. 2a and 2b), reflecting the local character of rainfall. Both ensembles show positive MSESS values for regions in Scandinavia and Eastern Europe, and to a lesser extent for Iberia and the British Isles (Fig. 2c and 2d). In CCLM_b1, predictive skill is also identified over Western Central Europe. Thus for CCLM_b1 positive skill is found for larger areas indicating an added value of the improved initialization procedure in baseline1 compared to baseline0.

Regarding wind speed, the predictive skill in CCLM_b0 (Fig. 2e) shows high MSESS values over Scandinavia, Iberia, southern Italy and along the coasts of the North and the Baltic Sea, while negative values are found e.g. over parts of France, southern Germany and the Alpine region. In CCLM_b1, the MSESS depicts positive values over most of Western and Central Europe, while negative values are now identified along the eastern coast of the Baltic Sea (Fig. 2f). Overall the predictive skill of CCLM_b0 is slightly higher and affects a larger area, indicating that the changes in the initialization method do not improve the results for wind speed.

We conclude that in terms of the MSESS accuracy there generally is a potential for skilful decadal predictions over Europe in the regional MiKlip ensembles. However, the skill pattern depends on the region and the variable. For individual regions, the initialisation of the hindcasts and decadal predictions lead to an added value for accurate (retrospective) forecasts several years ahead, while for some regions the uninitialized historical runs deliver more reliable predictions. Also the discrepancies between the two hindcast generations (CCLM_b0 and CCLM_b1) are rather heterogeneous. While for temperature we only found a slight shift in the pattern due to the different initialization methods, discrepancies can be large for precipitation and wind speed depending on the region.

A different picture is revealed when using the climatology as reference dataset for the MSESS computation (Fig. 3). Not surprisingly, for temperature the MSESS strongly increases to positive values for most of Europe (Fig. 3a and 3b). This is due to the strong positive trend in the observed temperature, which is predicted by the hindcasts but not captured by the climatology. Contrastingly, the MSESS with the climatology as reference generally decreases for wind speed in both CCLM_b0 and CCLM_b1 (Fig. 3e and 3f). The positive MSESS is maintained only for Northern Europe and CCLM_b0. Hence, the climatology is generally closer to the observations than both the hindcasts and the historical runs. We have analysed the respective spatial mean wind speed time series for the Iberian Peninsula (Prudence region 2) and CCLM_b0, where this effect is strongest. The wind speed shows a slight negative trend in both, CCLM_b0 and the historicals, while the trend is slightly positive for the observational dataset (not shown). At the same time, the decadal variability for wind speed is quite small over this region in all datasets (it ranges from 0.05 to -0.05 in the historicals, and from 0.02 to -0.02 in CCLM_b0 and E-OBS). Hence, the deviation of the climatology to the observations and thus its MSE are generally small in this region, resulting in a negative MSESS when using the climatology as reference (see also equation for MSESS in section 3.2). Rather robust results are found for precipitation, independently from the choice of the reference dataset for CCLM_b0 and CCLM_b1 (cf. Fig 3e and 3f with Fig. 2e and 2f).

For a better understanding of the skill scores and their relation, the different skill metrics are compared in scatter plots. Fig. 4a and 4c show scatter diagrams of CRPSS vs MSESS for temperature and wind speed, respectively, on individual grid point basis for CCLM_b0. Generally, the accuracy and the reliability can vary highly with geographical position. However, for the majority of the individual land grid points over Europe, positive MSESS are concurrent with positive CRPSS values for both displayed variables (upper right quadrant). Both skill scores are linked to each other, showing a quasi-linear dependency between CRPSS and MSESS. This is identified for both CCLM_b0 (Fig. 4a and 4c) and MPI_b0 (Fig. 4b and 4d). In particular, we found that positive values for CRPSS often accompany with a high accuracy of the decadal predictions. This is generally true for all variables and both ensembles considered here (not shown).

On the other hand, no such linear relationship is detected between ACC and MSESS (see Fig. 4e for wind speed in CCLM_b0). The ACC vs MSESS combination is clearly more strongly scattered than CRPSS vs MSESS, both in terms of the general spread and the peak values of the number of grid points with a given skill score combination. Hence, a low mean bias of decadal predictions (resulting in positive MSESS values) does not necessarily imply a realistic temporal evolution. Still, positive MSESS values correspond to positive ACC values for most of the individual grid points, indicating a high potential for skilful regional decadal predictions over Europe.

4.2 Does regional downscaling provide an added value for decadal predictions?

Recent studies document that the application of regional climate models may improve climate simulations, in particular over complex terrain (Berg et al., 2013; Feldmann et al., 2013; Hackenbruch et al., 2016). This is mainly due to a more realistic representation of the topography (e.g. mountain ranges or coast lines) in the RCMs compared to global-scale GCMs. In this section, we analyse whether the downscaling with a regional climate model also leads to an added value for decadal predictions over Europe.

Figure 4 indicates a shift of the overall distribution of skill scores towards higher values for the regionalised hindcasts compared to the global ones in the baseline0 ensemble. For temperature and wind speed, the core area of the skill values from the regional hindcasts (Fig 4a and 4c) is more confined to the upper right quadrant compared to the global ensemble (Fig 4b and 4d). This indicates an added value of downscaling for the accuracy as well as for the reliability. For the wind speed correlation the patterns are quite similar, as there is a clear shift towards an improved correlation and for a higher MSESS from the downscaling (Fig. 4e and 4f). In contrast, no added value on grid point scale is detected for precipitation in CCLM_b0, neither for the accuracy skill scores (MSESS and ACC) nor for CRPSS (not shown). With respect to the baseline1 ensemble, only the prediction skill for temperature is improved by the downscaling, while no or only a marginal added value of regionalization is found for precipitation and wind speed (not shown). Note that results may be different when spatial means over the Prudence regions are analysed (see below).

Ideally, an added value of downscaling should be accompanied by a positive absolute skill. Figure 5 depicts these two aspects for the three variables (2m temperature, precipitation, near-surface wind), the three verification metrics (MSESS, ACC, CRPSS) and additionally delta_ACC (see section 3.2), and the two ensemble generations (b0 and b1), as derived for the spatial means over the eight PRUDENCE regions (cf. Fig. 1). Green dots indicate higher skill scores for the CCLM ensembles compared to MPI-ESM-LR and thus an added value of downscaling. Red dots mean that the skill scores are lower in the CCLM ensembles, while in case of yellow dots the skill scores are equal in MPI-ESM-LR and CCLM. Red background color indicates a negative skill score and green color a positive skill for the respective metric. Thus, this figure can be interpreted along several dimensions: (i) the skill of the different climate variables (background color), (ii) the improvement by downscaling (dot color), (iii) the improvement from b0 to b1, (iv) the skill for different regions, (v) and the different skill metrics.

For temperature, CCLM_b1 mostly shows an added value compared to MPI_b1 as well as compared to CCLM_b0. For most regions, this is particularly expressed in all displayed metrics. For instance, with respect to the accuracy (MSESS, ACC, and delta_ACC) CCLM_b1 has higher skill in six or seven of eight PRUDENCE regions compared to MPI_b1. Only for the Alps (AL, Prudence region 6), no added value of downscaling is found in both ensemble generations for ACC and delta_ACC. Additionally, no benefit from downscaling could be detected for CCLM_b0 for the British Isles (BI – 1), France (FR – 3), Mid-Europe (ME - 4) and the Mediterranean Area (MD - 7), where CCLM_b1 performs better. In general, in CCLM_b1 there are more regions with positive skill scores in Southern Europe (IP - 2, AL - 6). In the Mediterranean region both ensemble generations depict only positive skill scores.

For precipitation, an improvement from downscaling is detected particularly for CCLM_b1 for the majority of metrics and regions. In addition, CCLM_b1 is clearly superior to CCLM_b0 with respect to both skill and added value. This indicates a positive effect of the improved initialization procedures in b1 compared to b0 (Pohlmann, 2013b). However, this improvement does not affect all regions. CCLM_b0 performs better than its successor for the Iberian Peninsula, whereas skill and/or added value are higher in CCLM_b1 for the regions in the North-West (BI, FR, ME) and North (SC). With respect to reliability, CCLM_b0 slightly outperforms CCLM_b1 for precipitation (CCLM_b0: seven regions with positive CRPSS, CCLM_b1: six regions with positive CRPSS), while for temperature a contrary result is found (three regions with positive CRPSS in CCLM_b0 and four in CCLM_b1).

For near surface wind, Fig. 5 shows heterogeneous results. CCLM_b0 has an added value of downscaling in more regions than CCLM_b1. Further, CCLM_b0 provides an added value for the CRPSS in eight regions, while for CCLM_b1 an improvement by downscaling is restricted to four regions with respect to the reliability. CCLM_b0 has a positive skill in the northern parts of the domain (BI, SC), whereas positive skill scores are found for CCLM_b1 over most other PRUDENCE regions at least for one skill metric.

The detected shift in the skill patterns between CCLM_b0 and CCLM_b1 can be expected due to the different initialization procedures of the two generations. However, there also seem to be regions with more stable skill properties for the CCLM ensembles: The Mediterranean area shows positive skill for all variables and metrics (except wind in CCLM_b0), though this high prediction skill is not necessarily accompanied by an improvement of the skill scores compared to the MPI-ESM-LR ensembles.

An added value of regionalization for the majority of variables and metrics can be found for Southern Europe (MD, IP) and Scandinavia. As these areas have complex coastlines and orography, this result may be indicative of a better representation of small-scale processes in the CCLM. On the other hand, for the Alps (AL) an added value of downscaling for all skill scores is only revealed for wind in CCLM_b0. The PRUDENCE region AL is the smallest of the regions, with the steepest orography. It might be that for the Alps an even higher resolution for the downscaling would be advantageous to improve the accuracy and reliability of the hindcasts.

We conclude that regional downscaling indeed may provide an added value for decadal predictions over Europe. However, while for some complex regions like MD, IP or SC this added value is to some extent systematic, for other areas in Europe the analysis reveals a mixed picture for the different variables and the skill metrics.

4.3 Does the regional decadal predictive skill depend on the ensemble size?

Past studies suggest that the ensemble size of a prediction system has an impact on the forecast skill of a model (Richardson, 2001; Ferro et al., 2008). Generally, there is consensus that the prediction skill for both seasonal and decadal predictions is enhanced when the number of ensemble members is increased. Kadow et al. (2014) analysed the global MiKlip baseline1 generation and concluded that the forecast accuracy for surface temperature for lead years 1 and 2-9 is improved for nearly the whole globe when the ensemble size is increased from 3 to 10 members. This is in line with the findings of Sienz et al. (2016), who examined the prediction skill for North Atlantic sea surface temperature in the same hindcast ensemble. Also for seasonal predictions of the North Atlantic Oscillation a forecast system profits from increasing size (e.g. Scaife et al., 2014). However, it is still open how a regional decadal forecast system does depend on the quantity of ensemble members. With this aim, we analysed the impact of the ensemble size on the predictive skill for the eight PRUDENCE regions in Europe in both the regional and the global MiKlip ensembles. In the following, results are only shown for the Iberian Peninsula (IP), as the findings are similar for the other PRUDENCE regions. Figure 6 exhibits the dependency of CRPSS, MSESS, and delta_ACC for lead years 1-5 (y-axis) on the ensemble size (x-axis) for all three variables spatially averaged over IP. For each ensemble size *n* (*n* varying between 2 and 10), the solid coloured lines depict the averaged skill scores for all permutations of *n*-member ensemble combinations for each of the four individual hindcast ensembles (MPI_b0, MPI_b1, CCLM_b0, and CCLM_b1). Ranked probability skill scores like the CRPSS may be negatively biased for small ensembles sizes (e.g. Ahrens and Walser, 2008), and different bias correction methods exist to overcome this issue. However, as this correction would only affect the CRPSS, a direct comparability with the MSESS and the ACC would not be warranted anymore after such a correction. To ensure the comparability of the results for the three skill metrics we therefore decided not to use a de-biased version of the CRPSS in this study.

Enhanced predictive skill can be observed when the number of members is stepwise increased for both the global and the regional hindcast ensembles. MSESS and CRPSS show a rather logarithmic relationship with increasing *n*, depicting the highest skill scores for the 10 member ensembles for all three variables (Figure 6a-c and 6g-i). On the other hand, the lowest skill scores (often with negative values for CRPSS and MSESS) are always found for the 2-member ensembles. This ensemble size dependency of MSESS and CRPSS is systematic and is detected in both hindcast generations for all variables over all eight PRUDENCE regions (not shown), regardless whether the skill scores are negative or positive. In some cases, the ensemble size increase even leads to a shift from negative MSESS and CRPSS values to positive values in one or more of the ensembles (e.g. Fig. 6a, 6g, and 6i). In contrast, no systematic conclusion can be stated for the delta_ACC, as the ensemble size dependency of the predictive skill depends on the variable and the considered MiKlip ensemble (Fig. 6d-f). But even here a larger ensemble size is advantageous, as negative skill scores become more robust (cf. MPI_b0 in Fig. 6f). Nevertheless, there are also examples for delta_ACC where the ensemble size dependency is similar to that of MSESS and CRPSS, like e.g. for temperature (Fig. 6d). These results suggest that a decadal prediction system generally benefits from larger ensemble sizes, either in terms of more skilful and reliable decadal forecasts or at least of a reduction of the bias or the uncertainty, depending on the variable and the hindcast generation. Note that for most variables and skill scores the hindcast generation is more important for the skill than the resolution. In addition, most diagrams indicate an added value of downscaling. For temperature and wind speed, both generations of CCLM surpass their MPI counterparts for all skill scores, indicating a systematic added value of downscaling. This is particularly visible for wind in the b0 ensemble, where the prediction skill of CCLM is distinctly better than for MPI-ESM-LR (Fig. 6c, 6f, and 6i). This is mainly due to higher skill scores over orographic structured terrains of IP in CLM_b0 compared to MPI_b0 (not shown).

For ensembles with less than 10 members, the skill scores of all possible *n*-member ensemble combinations are averaged. This is illustrated for the MSESS for precipitation in the CCLM_b0 ensemble (see box-whisker plots in Fig. 6b). Given that we are doing permutations without replacement, the spread between the individual *n*-member ensembles declines with an increasing number of members *n*, and this decline should therefore not be over-interpreted. Nevertheless, the spread is quite large not only for small ensemble sizes but also for ensembles with *n*>5: For instance, the MSESS varies between -2.6 and +0.7 for the 2-member ensembles (Fig. 6b), and even for the 7-member ensemble quite different results can be found depending on the selection of the ensemble members, ranging from high positive MSESS values to zero. These results clearly demonstrate the necessity of using large ensembles to reduce uncertainties.

We conclude that the predictive skill with respect to both accuracy and model spread is generally improved when the size of the hindcast ensembles increases. This is valid for all variables, regions, and hindcast ensembles considered in this study. The skill scores converge towards a certain value in most cases for MSESS and CRPSS in all hindcasts (see Fig. 6a-c and 6g-i). The increments in added value by increasing the number of ensemble members decrease for more than 5 members. Nevertheless, it is recommended to use ten members or more for the skill assessment of decadal predictions on the regional scale.

4.4 How does the sample size affect the skill estimates?

A lesson learned from the CMIP5 decadal experiments is that more starting years and thus a larger sample size is beneficial to establish robust skill estimates (Boer et al., 2016). This has been reflected in the progress from the first global MiKlip hindcast generation baseline0 to the second generation baseline1. Whereas baseline0 provides ten ensemble members every fifth year (compliant with the CMIP5 experimental protocol), baseline1 provides this ensemble size for each starting year of the hindcast period. To assess the impact of the small sample size with five starting years (used elsewhere in the paper) on the robustness of our main conclusions we performed a sensitivity analysis with the global baseline1 ensemble, for which the largest sample is available. For this, we compared the sample with ten-yearly starting dates with the full yearly initialized MPI-ESM-LR baseline1 ensemble over the same period from 1960 to 2000.

Fig. 7 presents a comparison between the ACC scores for the small (left; 5 starting years) and the large sample size (right; 41 starting years). For all three variables. the score maps show in general comparable spatial distributions. The skill maps for the larger sample size usually depict a smoother spatial distribution with less extreme skill values. The regional averages over most of the PRUDENCE regions are comparable. However, in some regions larger differences can occur: For temperature over Ireland and Scotland, for precipitation over parts of France and Eastern Europe and for wind from north-eastern Spain towards the Alps. Similar results are found for MSESS (not shown), for which not only the sample of MPI_b1 is increased but also of the uninitialized historical runs.

It is obvious that a larger sample size increases the robustness of the skill assessment, especially with respect to quantitative estimates. Therefore, this work supports the recommendations made for CMIP6 by Boer et al. (2016) to generate hindcast ensembles with yearly starting dates. Nevertheless, using the smaller sample size already represents the general features of the regional distribution. Therefore, the qualitative findings from chapter 4.1 are confirmed by the analysis of the larger sample size. The results regarding the added value and the ensemble size dependence are less affected by the sample size. Given the above findings, we conclude that the results obtained here for a limited sample size are qualitatively comparable to those which would be obtained for a larger sample size.

5. Summary and discussion

In this study, the decadal predictability in the regional MiKlip decadal prediction system is analysed for temperature, precipitation, and wind speed over Europe and compared to the forecast skill of the global ensemble. The goal is to assess the prospect of such a system for the application in forecasts on decadal timescales. Focus is given to years 1-5 after initialization. Three skill scores are used to quantify the accuracy and the reliability of the two different MiKlip hindcast generations. The main findings of our study can be summarized as follows:

- There is a potential for regional decadal predictability over Europe for temperature, precipitation, and wind speed in the MiKlip system, but the predictive skill depends on the variable, the region, and the hindcast generation.

- The MiKlip prediction system may distinctly benefit from regional downscaling. An added value in terms of accuracy and reliability is particularly revealed for temperature over the British Isles (BI), Scandinavia (SC), the Iberian Peninsula (IP), and for precipitation over the British Isles (BI), Scandinavia (SC), Mid-Europe (ME), and France (FR) for the b1 generation. Most of these regions are characterized by complex coastlines and orography, which indicates that the better representation of topographic structures in the regionalised hindcasts may improve the predictive skill.
- The improvement of the initialization procedure from baseline0 to baseline1 as described in Pohlmann et al. (2013b) increases the overall predictive skill in the downscaled MiKlip hindcasts over Europe, at least for precipitation and temperature. But improvement of the skill varies between variable and region. The skill for temperature increases around the Mediterranean Sea and parts of Scandinavia from b0 to b1. For precipitation the skill of b1 compared to b0 is higher in all regions but the Iberian Peninsula and Eastern Europe. Only for wind speed there is mostly no benefit from the improved initialization.
- A systematic enhancement of MSESS and CRPSS skill scores is found with increasing ensemble size, and a number of 10 members is found to be suitable for decadal predictions. This is valid for all variables and European regions in the global and regional MiKlip ensembles.
- As tested for the MPI_b1 data, which offer a full ten member ensemble for each starting year, a larger sample size would lead to similar results as presented here, Nevertheless, such an increase would improve the robustness of the skill maps.

Müller et al. (2012) and Pohlmann et al. (2013b) had found systematic prediction skills for surface temperature over large parts of the North-Atlantic and Europe in both global generations (baseline0, baseline1). From the results of our study, it is apparent that the Mediterranean Area and the Iberian Peninsula seem to be key European regions for decadal predictability with the regional prediction system. This is in line with findings from Guemas et al. (2015) and may be related to skilful predictions of the AMO (Garcia-Serrano et al., 2012; Guemas et al., 2015). Due to the rather non-linear relationship of these large-scale North Atlantic features to regional atmospheric conditions over Europe, the mechanisms steering the decadal variability and predictability of climate variables in European regions are thus more complex. The decadal variability of regional precipitation, temperature, and wind speed over most parts of Europe is largely affected by the North Atlantic oscillation, but its skilful decadal predictability over the continent is still under debate. With this respect, a better understanding of the mechanisms relevant for the regional climate over Europe on the decadal time scale is required, as was for example obtained for the tropical Atlantic (Dunstone et al., 2011). This is an objective of the ongoing second phase of the MiKlip project.

The skill scores may strongly vary between neighbouring grid points. Comparable results were found by e.g. Guemas et al. (2015), who detected a rather diffuse pattern for the accuracy of decadal predictions over Europe for seasonal temperature and precipitation. This might at least partly be due to spatial and temporal inhomogeneity of the gridded observational references. A more realistic assessment of the prediction skill can be made by considering spatial means (Goddard et al.,

2013) which was mostly considered in this study. In line with e.g. Kadow et al. (2016), we could show that an enlargement of the ensemble size up to 10 members results in an improvement of the prediction skill over Europe. However, prediction skill could further benefit from even larger ensemble sizes, especially in areas with low signal-to-noise ratio (cf. Sienz et al., 2016).

Bias and drift adjustment (e.g., Boer et al., 2016) provides prospect in skill improvement not only for GCMs but also for RCMs. This is particularly the case for ensemble simulations run with full-field initialization (like the third MiKlip generation prototype, not analysed here; cf. Marotzke et al., 2016). While bias and drift adjustment methods have improved the forecast skill of near-term climate prediction (e.g., Kruschke et al., 2016), such corrections are less important for the baseline0 and baseline1 ensembles analysed here as they were generated with anomaly initialisation (Marotzke et al., 2016). Nevertheless, bias correction and calibration are an important topic in the second phase of MiKlip.

Due to the high computational costs of dynamical downscaling, only five starting dates (one per decade) are available for the regional MiKlip ensemble (see section 2). This is a shortcoming regarding the statistical significance of the results and some of the statements presented in this study. However, we could show that the qualitative findings are only partly influenced by the limited number of available hindcasts and that the main conclusions can be regarded as robust. The statistical significance will be easier to quantify when the regional simulations for the newest Miklip ensemble generation are available with annual starting dates over more than 50 years. On the other hand, regional decadal forecasts may have advantages beyond the examples discussed in this paper. For example, RCMs enable the integration of improved components of the hydrological cycle or climate-system components with memory on multi-year time-scales like soil moisture (Khodaya et al., 2014; Sein et al., 2015). Kothe et al. (2016) has shown that extracting the initial state of the deep soil in the RCMs from regional data assimilation schemes may improve decadal predictions. Further, Akhtar et al. (2017) demonstrated that the regional feedback between large water bodies and the atmosphere play a major in the regional climate system. This feedback can only be captured in regionalized climate predictions by a dynamic RCM-ocean coupling. Most of the approaches mentioned above are ongoing within the second phase of MiKlip and are expected to enhance the decadal predictability over Europe. We thus conclude that a decadal prediction system would clearly benefit from a regional forecast ensemble.

The regional decadal prediction system generated by the MiKlip consortium comprises altogether 1000 years (two hindcast generations, each of them comprising ten hindcast members for five starting years) of simulations with 0.22° for the entire EURO-CORDEX region, which is a to our best knowledge unprecedented. Hence, this ensemble enabled us to gain important insights into different aspects and the prospects of regional downscaling for decadal predictions, and serve as a good basis for future studies. In the ongoing second phase of MiKlip it is planned to downscale a complete ensemble hindcast generation with ten members for more than 50 starting years, giving altogether more than 5000 years.

Author Contributions

MR, HF, SM and MU developed the concept of the paper; MR, HF and JGP wrote the first manuscript draft. MR, HF, SM, MU, NL and JM contributed with data analysis and analysis tools. HF, SM, MR, BA, BF contributed with RCM simulations. MK and WM contributed with the global MPI-ESM-LR simulations and prepared boundary conditions for RCM simulations. CK leads the MiKlip-C consortium, with CO-Is BA, BF, JGP, GS. All authors contributed with ideas, interpretation of the results and manuscript revisions.

Acknowledgments

MiKlip is funded by the German Federal Ministry for Education and Research (BMBF, contracts: 01LP1518 A-D and 01LP1519) All simulations were carried out at the German Climate Computing Centre (DKRZ), which also provided all major data services. We acknowledge the E-OBS data set from the EU-FP6 project ENSEMBLES (http://ensembles-eu.metoffice.com) and the data providers in the ECA&D project (http://www.ecad.eu). We thank the European Centre for Medium-Range Weather Forecasts (ECMWF) for their ERA-40 and ERA-Interim Reanalysis data (http://apps.ecmwf.int/datasets/). JGP thanks the AXA Research Fund for support. We thank past and present members of the MiKlip –C (Regionalization) group for discussions and comments. We thank the Reviewers for their comments, which helped to improve the manuscript.

References

Akhtar, N., Brauch, J., and Ahrens, B.: Climate Modeling over the Mediterranean Sea: Impact of Resolution and Ocean Coupling, Clim. Dynam., doi:10/1007/s00382-017-3570-8, 2017.

Balmaseda, M. A., Mogensen, K., and Weaver, A. T.: Evaluation of the ECMWF ocean reanalysis system ORAS4, Q. J. R. Meteor. Soc., 139, 1132-1161., doi:10.1002/qj.2063, 2013.

Benestad, R. E. and Mezghani, A.: On downscaling probabilities for heavy 24-hour precipitation events at seasonal-to-decadal scales, Tellus A, 67, 25954, doi:10.3402/tellusa.v67.25954, 2015.

Berg, P., Wagner, S., Kunstmann, S., and G. Schaedler: High resolution regional climate model simulations for Germany: part I – validation, Clim. Dynam., 40, 401-414, 2013.

Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., Kushnir, Y., Kimoto, M., Meehl, G. A., Msadek, R., Mueller, W. A., Taylor, K. E., Zwiers, F., Rixen, M., Ruprich-Robert, Y., and Eade, R.: The Decadal Climate Prediction Project (DCPP) contribution to CMIP6, Geosci. Model Dev., 9, 3751-3777, doi:10.5194/gmd-9-3751-2016, 2016.

Chikamoto Y., Kimoto, M., Ishii, M., Mochizuki, T., Sakamoto, T. T., Tatebe, H., Komuro, Y., Watanabe, M., Nozawa, T., Shiogama, H., Mori, M., Yasunaka, S., and Imada, Y.: An overview of decadal climate predictability in a multi-model ensemble by climate model MIROC, Clim. Dynam., 40, 1201-1222, doi:10.1007/s00382-012-1351-y, 2012.

Christensen, J.H. and Christensen, O.B.: A summary of the PRUDENCE model projections of changes in European climate by the end of this century, Climate Change, 81, 7-30, doi:10.1007/s10584-006-9210-7, 2007.

Corti S., Palmer, T., Balmaseda, M., Weisheimer, A., Drijfhout, S., Dunstone, N., Hazeleger, W., Kröger, J., Pohlmann, H., Smith, D., von Storch. J.-S., and Wouters, B.: Impact of Initial Conditions versus External Forcing in Decadal Climate Predictions: A Sensitivity Experiment, J. Climate, 28, 4454–4470, doi:10.1175/JCLI-D-14-00671.1, 2015.

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

Doblas-Reyes, F. J., Andreu-Burillo, I., Chikamoto, Y., García-Serrano, J., Guemas, V., Kimoto, M., Mochizuki, T., Rodrigues, L. R. L. and van Oldenborgh, G. J.: Initialized near-term regional climate change prediction, Nature Commun., 4, 1715, doi:10.1038/ncomms2704, 2013.

Dunstone, N. J., Smith, D. M., and Eade, R.: Multi-year predictabilityof the tropical Atlantic atmosphere driven by the high latitude North Atlantic Ocean, Geophys. Res. Lett., 38, L14701, doi:10.1029/2011GL047949, 2011.

Feldmann, H., Schaedler, G., Panitz, H.-J., and Kottmeier, C.: Near future changes of extreme precipitation over complex terrain in Central Europe derived from high resolution RCM ensemble simulations, Int. J. Climatol., 33, 1964-1977, 2013.

Ferro, C. A. T., Richardson, D. S., and Weigel, A. P.: On the effect of ensemble size on the discrete and continuous ranked probability scores, Meteorol. Appl., 15, 1, 19-24, doi:10.1002/met.45, 2008.

Garcia-Serrano, J., Doblas-Reyes, F. J., and Coelho, C. A. S.: Understanding Atlantic multi-decadal variability prediction skill, Geophys. Res. Lett., 39, L18708, doi:10.1029/2012GL053283, 2012.

Giorgetta, M. A., Jungclaus, J. J., Reick, C. H., Legutke, S., Bader, J., Böttinger, M. and Brovkin, V., Crueger, T., Esch, M., Fieg, K., Glushak, K., Gayler, V., Haak., H., Hollweg, H.-D., Ilyina, T., Kinne, S., Kornblueh, L., Matei, D., Mauritsen, T., Mikolajewicz., U., Mueller, W. A., Notz, D., Pithan, F., Raddatz, T., Rast, S., Redler, R., Roeckner, E., Schmidt, H., Schnur, R., Segschneider, J., Six, K. D., Stockhause, M., Timmreck, C., Wegner, J., Widmann, H., Wieners, K.-H., Claussen, M., Marotzke, J., and Stevens, B.: Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5, J. Adv. Model. Earth Sy., 5, 572–597, doi:10.1002/jame.20038, 2013.

Giorgi, F., Jones, C., and Asrar, G. R.: Addressing climate information needs at the regional level: the CORDEX framework, Bulletin of the World Meteorological Organization, 58, 175-183, 2006.

Goddard, L., Kumar, A., Solomon, A., Smith, D., Boer, G., Gonzalez, P., Kharin, V., Merryfield, W., Deser, C., Mason, S. J., Kirtman, B. P., Msadek, R., Sutton, R., Hawkins, E., Fricker, T., Hegerl, G., Ferro, C. A. T., Stephenson, D. B., Meehl, G. A., Stockdale, T., Burgman, R., Greene, A. M., Kushnir, Y., Newman, M., Carton, J., Fukumori, I., and Delworth, T.: A verification framework for interannual-to-decadal predictions experiments, Clim. Dynam., 40, 245-272, doi:10.1007/s00382-012-1481-2, 2013.

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

Hackenbruch, J., Schaedler, G., and Schipper, J. W.: Added value of high-resolution regional climate simulations for regional impact studies, Meteorol. Z., 25, 291-304, doi:10.1127/metz/2016/0701, 2016.

Haas, R., Reyers, M., and Pinto, J. G.: Decadal predictability of regional-scale peak winds over Europe based on MPI-ESM-LR, Meteorol. Z., 25, 739-752, doi:10.1127/metz/2015/0583, 2016.

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

Ho, C. K., Hawkins, E., Shaffrey, L., Bröcker, J., Hermanson, L., Murphy, J. M., Smith, D. M., and Eade, R.: Examining reliability of seasonal to decadal sea surface temperature forecasts: The role of ensemble dispersion, Geophys. Res. Lett., 40, 5770-5775, doi:10.1002/2013GL057630, 2013.

Jungclaus, J. H., Fischer, N., Haak, H., Lohmann, K., Marotzke, J., Matei, D., Mikolajewicz, U., Notz, D., and von Storch, J.-S.: Characteristics of the ocean simulations in MPIOM, the ocean component of the MPI-Earth system model, J. Adv. Model. Earth Sy., 5, 422-446, doi:10.1002/jame.20023, 2013.

Kadow, C., Illing, S., Kunst, O., Rust, H. W., Pohlmann, H., Müller, W. A., and Cubasch, U.: Evaluation of forecasts by accuracy and spread in the MiKlip decadal climate prediction system, Meteorol. Z., 25, 631-643, doi:10.1127/metz/2015/0639, 2016.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C. Wang, J., Jenne, R. and Joseph, D.: The NCEP/NCAR 40-Year Reanalysis Project, B. Am. Meteorol. Soc., 77, 437-471, doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2, 1996.

Khodayar, S., Selinger, A., Feldmann, H., Kottmeier, Ch.: 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, 2014.

Kothe, S., Tödter, J., and Ahrens, B.: Strategies for soil initialisation in regional decadal climate predictions, Meteorol. Z., 25, 775-794, doi:10.1127/metz/2016/0729, 2016.

Kröger, J., Müller, W. A., and von Storch, J.-S.: Impact of different ocean reanalyses on decadal climate prediction, Clim. Dynam., doi:10.1007/s00382-012-1310-7, 2012.

Kruschke, T., Rust, H. W., Kadow, C., Leckebusch, G. C., and Ulbrich, U.: Evaluating decadal predictions of northern hemispheric cyclone frequencies, Tellus A, 66, 22830, doi:10.3402/tellusa.v66.22830, 2014.

Kruschke, T., Rust, H. W., Kadow, C., Müller, W. A., Pohlmann, H., Leckebusch, G. C., and Ulbrich, U.: Probabilistic evaluation of decadal prediction skill regarding Northern Hemisphere winter storms, Meteorol. Z., 25, 721-738, doi:10.1127/metz/2015/0641, 2016.

Li, H., Ilyina, T., Müller, W. A., and Sienz, F.: Decadal predictions of the North Atlantic CO2 uptake, Nature Commun., 7, doi:10.1038/ncomms11076, 2016.

Marotzke J., Müller, W. A., Vamborg, F. S. E., Becker, P., Cubasch, U., Feldmann, H., Kaspar, F., Kottmeier, C., Marini, C., Polkova, I., Prömmel, K., Rust, H. W., Rust, H. W., Stammer, D., Ulbrich, U., Kadow, C., Köhl, A., Kröger, J., Kruschke, T., Pinto, J. G., Pohlmann, H., Reyers, M., Schröder, M., Sienz, F., Timmreck, C., and Ziese, M.: MiKlip – a National Research Project on Decadal Climate Prediction, B. Am. Meteorol. Soc., Early Online Releases, doi:10.1175/BAMS-D-15-00184.1, 2016.

Matei, D., Pohlmann, H., Jungclaus, J. H., Müller, W. A., Haak, H., and Marotzke, J.: Two tales of initializing decadal climate prediction experiments with the ECHAM5/MPI-OM model, J. Climate, 8502-8523, doi:10.1175/JCLI-D-11-00633.1, 2012.

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., Dixon, K., Giorgetta, M. A., Greene, A. M., Hawkins, E., Hegerl, G., Karoly, D., Keenlyside, N. S., Kimoto, M., Kirtman, B., Navarra, A., Pulwarty, R., Smith, D., Stammer, D., and Stockdale, T.: Decadal Prediction, B. Am. Meteorol. Soc., 90, 1467-1485, doi:10.1175/2009BAMS2778.1, 2009.

Meehl, G. A., Goddard, L., Boer, G., Burgman, R., Branstator, G., Cassou, C., Corti S., Danabasoglu, G., Doblas-Reyes, F., Hawkins, E., Karspeck, A., Kimoto, M., Kumar, A., Matei, D., Mignot, J., Msadek, R., Navarra, A., Pohlmann, H., Rienecker, M., Rosati, T., Schneider, E., Smith, D., Sutton, R., Teng, H., van Oldenborgh, G. J., Vecchi, G., and Yeager, S.: Decadal Climate Prediction: An Update from the Trenches, B. Am. Meteorol. Soc., 95, 243–267, doi:10.1175/BAMS-D-12-00241.1, 2014.

Mieruch, S., Feldmann, H., Schädler, G., Lenz, C.-J., Kothe, S., and Kottmeier, C.: 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, 2014.

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

Müller, W. A., Baehr, J., Haak, H., Jungclaus, J. H., Kröger, J., Matei, D., Notz, D., Pohlmann, H., von Storch, J.-S., and Marotzke, J.: Forecast skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology, Geophys. Res. Lett., 39, L22707, doi:10.1029/2012GL053326, 2012.

Pohlmann, H., Smith, D. M., Balmaseda, M. A., Keenlyside, N. S., Masina, S., Matei, D., Müller, W. A., and P. Rogel, P.: Predictability of the mid-latitude Atlantic meridional overturning circulation in a multi-model system, Clim. Dynam., 41, 775-785, doi:10.1007/s00382-013-1663-6, 2013a.

Pohlmann H., Müller, W. A., Kulkarni, K., Kameswarrao, M., Matei, D., Vamborg, F. S. E., Kadow, C., Illing, S., and Marotzke, J.: Improved forecast skill in the tropics in the new MiKlip decadal climate predictions, Geophys. Res. Lett., 40, 5798–5802, doi:10.1002/2013GL058051, 2013b.

Richardson, D.S.: Measures of skill and value of ensemble predictions systems, their interrelationship and the effect of ensemble size, Q. J. R. Meteor. Soc., 1277, 2473-2489, doi:10.1002/qj.49712757715, 2001.

Robson, J., Sutton, R., and D. Smith: Predictable climate impacts of the decadal changes in the ocean in the 1990s, J. Climate, doi:10.1175/JCLI-D-12-00827.1, 2013.

Rockel, B., Will, A., and A. Hense: The Regional Climate Model COSMO-CLM (CCLM), Meteorol. Z., 17, 347- 348, doi:10.1127/0941-2948/2008/0309, 2008.

Scaife, A. A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R. T., Dunstone, N., Eade, R., Fereday, D., Folland, C. K., Gordon, M., Hermanson, L., Knight, J. R., Lea, D. J., MacLachlan, C., Maidens, A., Martin, M., Peterson, A. K., Smith, D., Vellinga, M., Wallace, E., Waters, J., and Williams, A.: Skillful long-range prediction of European and North American Winters, Geophys. Res. Lett., 41, 2514-2519, doi:10.1002/2014GL059637, 2014.

Sein, D. V., Mikolajewicz, U., Gröger, M., Fast, I., Cabos, W., Pinto, J. G., Hagemann, S., Semmler, T., Izquierdo, A., and Jacob, D.: Regionally coupled atmosphere - ocean – sea ice – marine biogeochemistry model ROM: 1. Description and validation, J. Adv. Model. Earth Sy., 7, 268–304, doi:10.1002/2014MS000357, 2015.

Sienz, F., Müller, W. A., and Pohlmann, H.: Ensemble size impact on the decadal predictive skill assessment, Meteorol. Z., 25, 6, 645–655, 2016.

Stevens, B., Giorgetta, M. A., Esch, M., Mauritsen, T., Crueger, T., Rast, S., Salzmann, M., Schmidt, H., Bader, J., Block, K., Brokopf, R., Fast, I., Kinne, S., Kornblueh, L., Lohmann, U., Pincus, R., Reichler, T., and Roeckner, E.: Atmospheric component of the MPI-M Earth System Model: ECHAM6, J. Adv. Model. Earth Sy., 5, 146-172, doi:10.1002/jame.20015, 2013.

Sutton, R. T., and Dong, B.: Atlantic Ocean influence on a shift in European climate in the 1990s, Nature Geosc., 5, 788-792, doi:10.1038/NGEO1595, 2012.

Sutton, R.T. and Hodson, D.L.R: Atlantic Ocean Forcing of North American and European Summer Climate, Science, 309, 5731, 115-118, doi:10.1126/science.1109496, 2005.

Taylor, K.E., Stouffer, R.J., and Meehl, G.A.: An Overview of CMIP5 and the Experiment Design, B. Am. Meteorol. Soc., 93, 485–498, doi:10.1175/BAMS-D-11-00094.1, 2012.

Uppala, S. M., KÅllberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V. D. C., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Berg, L. Van De., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., Mcnally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J.: The ERA-40 re-analysis. Q. J. R. Meteor. Soc., 131, 2961–3012, doi:10.1256/qj.04.176, 2005.

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

Yeager, S., Karspeck, A., Danabasoglu, G., Tribbia, J., and Teng, H.: A decadal prediction case study: Late twentieth-century North Atlantic Ocean heat content, J. Climate, 25, 5173-5189, doi:10.1175/JCLI-D-11-00595.1, 2012.

Figures

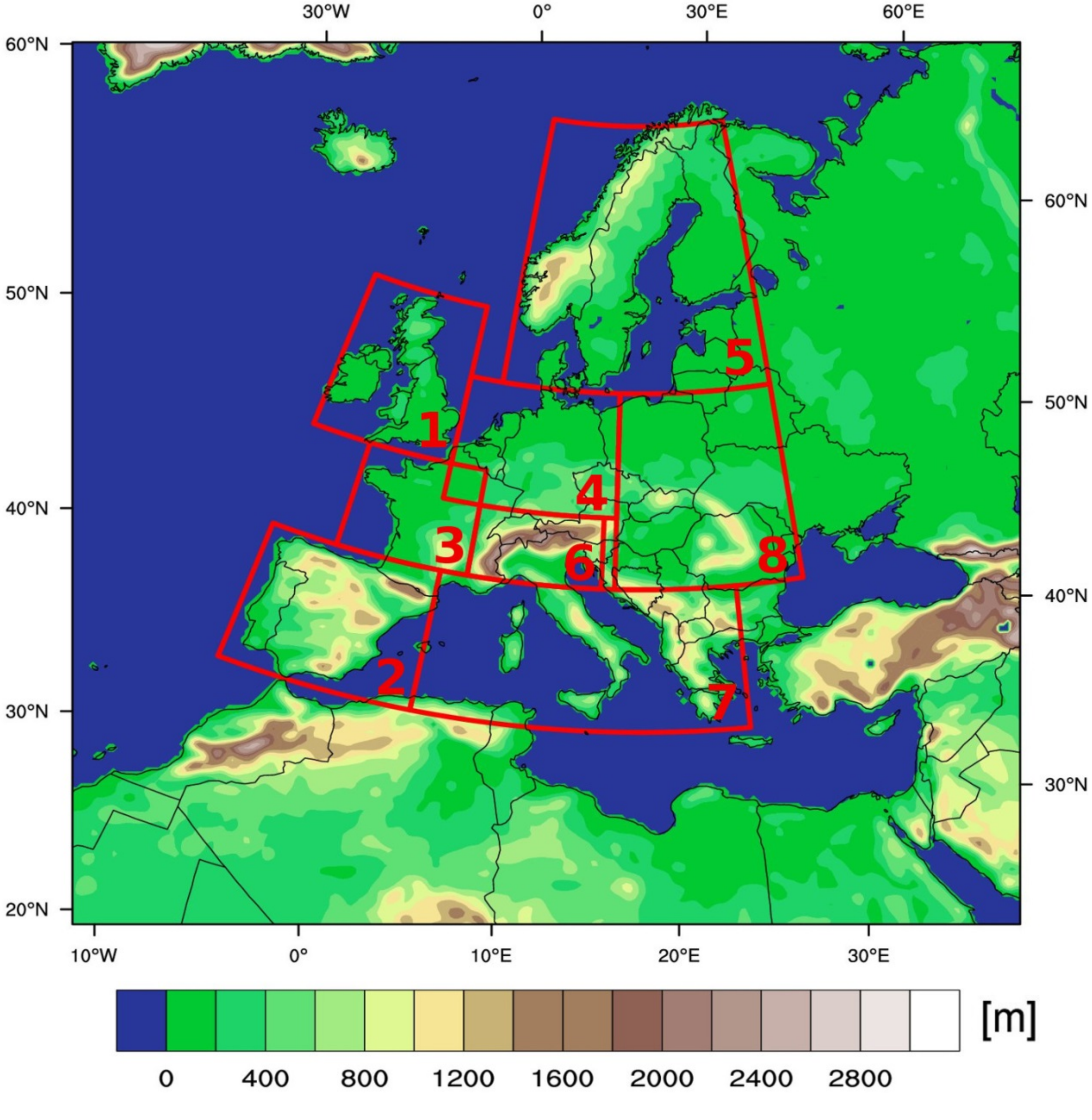


Figure 1: CCLM modelling domain (= EURO-CORDEX domain): Modell orography and PRUDENCE regions. 1: British Isles BI; 5 2: Iberian Peninsula IP; 3: France FR; 4: Mid-Europe ME; 5: Scandinavia SC; 6: Alps AL; 7: Mediterranean MD; 8: Eastern Europe EA.

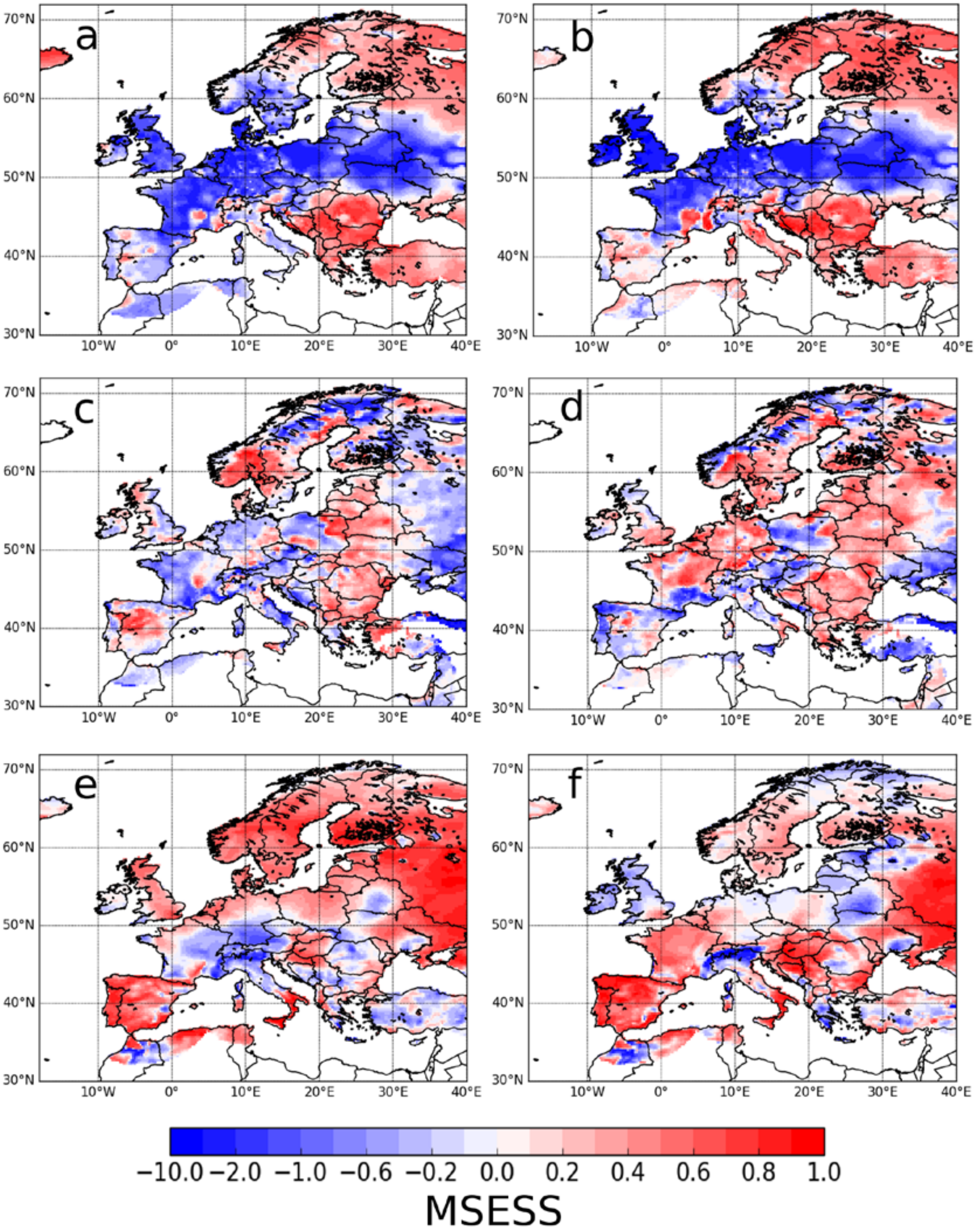


Figure 2: Spatial distribution of the MSESS for the multi-annual mean anomalies of lead years 1-5 for (a) temperature in CCLM_b0, (b) temperature in CCLM_b1, (c) precipitation in CCLM_b0, (d) precipitation in CCLM_b1, (e) wind speed in CCLM_b0, and (f) wind speed in CCLM_b1. As reference dataset we have used the uninitialized historical ensemble.

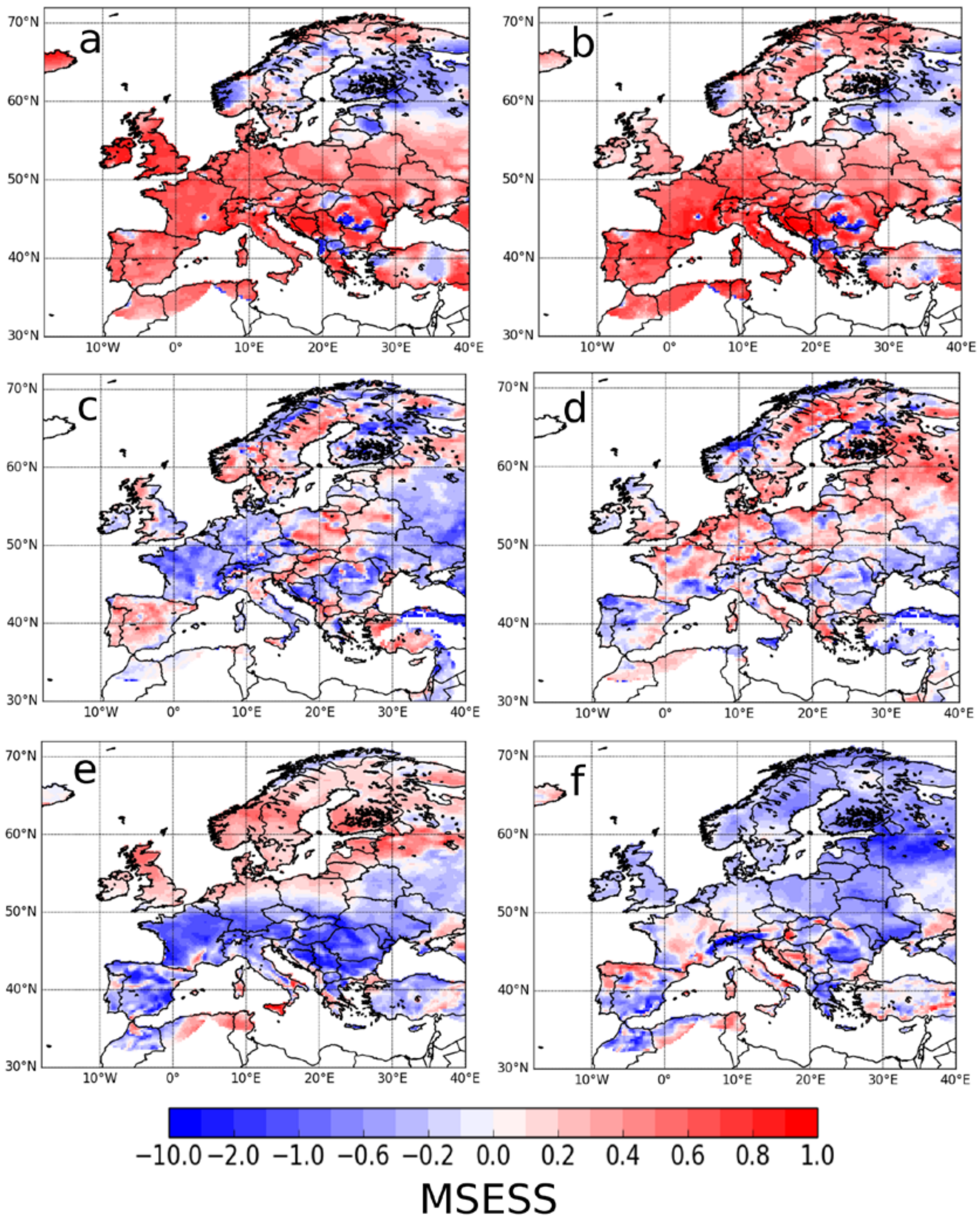


Figure 3: **Spatial distribution of the MSESS for the multi-annual mean anomalies of lead years 1-5 for (a) temperature in CCLM_b0, (b) temperature in CCLM_b1, (c) precipitation in CCLM_b0, (d) precipitation in CCLM_b1, (e) wind speed in CCLM_b0, and (f) wind speed in CCLM_b1. As reference dataset we have used the climatology.**

5

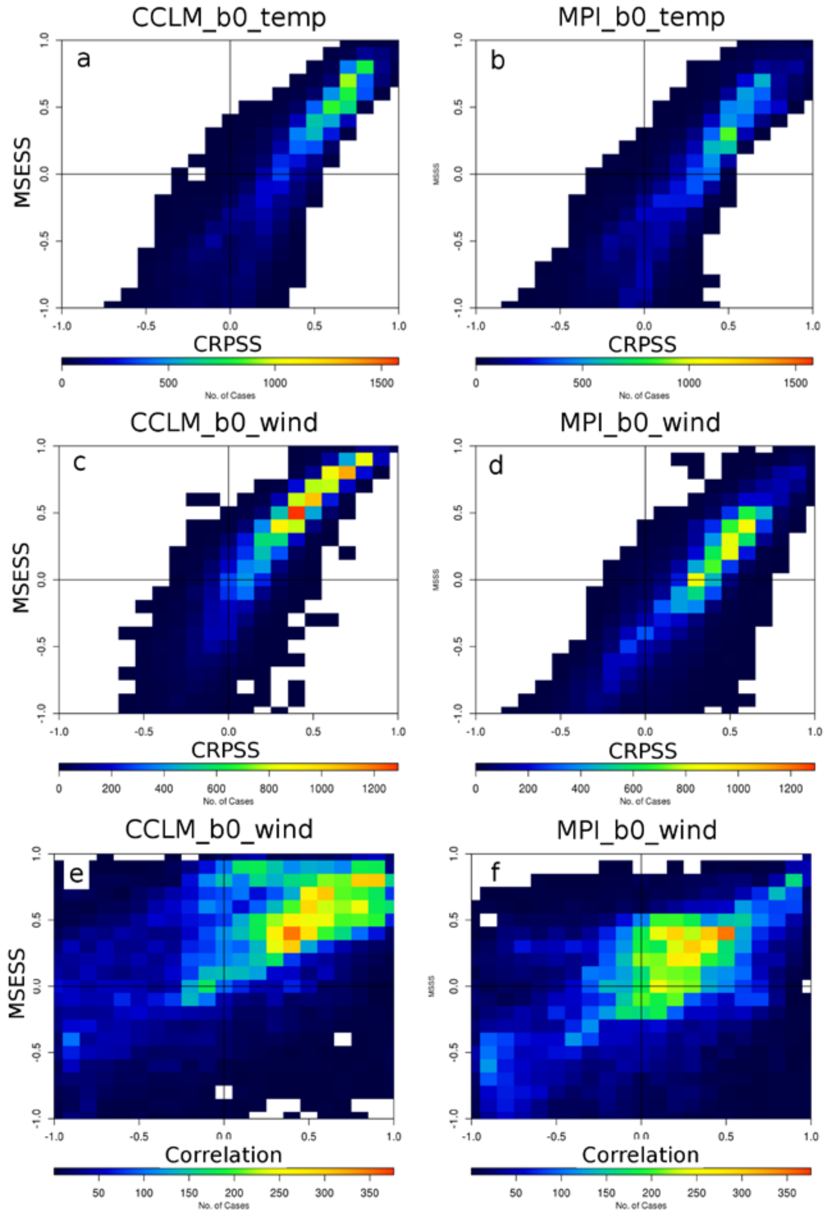
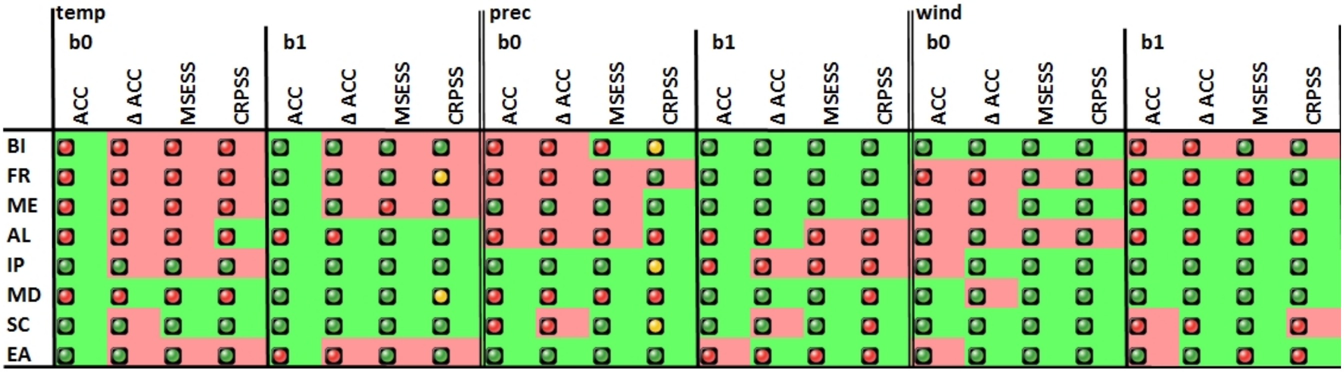


Figure 4: Scatter diagrams for CRPSS (x-axis) vs MSESS (y-axis) for temperature at all individual EURO-CORDEX grid points for the multi-annual mean anomalies of lead years 1-5 in (a) CCLM_b0 and (b) MPI_b0. (c), (d) as (a), (b) but for CRPSS vs MSESS for wind. (e), (f) as (a), (b) but for ACC vs MSESS for wind. Colours denote the number of grid points over Europe with a given skill score combination. For MSESS and CRPSS we have used the uninitialized historical ensemble as reference dataset. Note the different scaling of the colour bars.



	temp								prec								wind							
	b0				b1				b0				b1				b0				b1			
	ACC	Δ ACC	MSESS	CRPSS	ACC	Δ ACC	MSESS	CRPSS	ACC	Δ ACC	MSESS	CRPSS	ACC	Δ ACC	MSESS	CRPSS	ACC	Δ ACC	MSESS	CRPSS	ACC	Δ ACC	MSESS	CRPSS
BI	🟩🔴	🟥🔴	🟥🔴	🟥🔴	🟩🟢	🟥🟢	🟥🟢	🟥🟢	🟥🔴	🟥🔴	🟩🔴	🟩🟡	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟥🔴	🟥🔴	🟥🟢	🟥🟢
FR	🟩🔴	🟥🔴	🟥🔴	🟥🔴	🟩🟢	🟥🟢	🟥🟢	🟥🟡	🟥🔴	🟥🔴	🟥🟢	🟥🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟥🔴	🟥🔴	🟥🟢	🟥🟢	🟩🔴	🟩🔴	🟩🔴	🟩🟢
ME	🟩🔴	🟥🔴	🟥🔴	🟥🔴	🟩🟢	🟥🟢	🟥🔴	🟥🟢	🟥🟢	🟥🟢	🟥🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟥🟢	🟥🟢	🟩🟢	🟩🟢	🟩🔴	🟩🔴	🟩🔴	🟩🔴
AL	🟩🔴	🟥🔴	🟥🔴	🟩🔴	🟩🔴	🟩🔴	🟩🟢	🟩🟢	🟥🔴	🟥🔴	🟥🔴	🟩🔴	🟩🔴	🟩🔴	🟥🔴	🟥🔴	🟥🟢	🟥🟢	🟥🟢	🟥🟢	🟩🔴	🟩🔴	🟩🔴	🟩🔴
IP	🟩🟢	🟥🟢	🟥🟢	🟥🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟡	🟩🔴	🟥🔴	🟥🔴	🟥🔴	🟥🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢
MD	🟩🔴	🟩🔴	🟩🔴	🟩🔴	🟩🟢	🟩🟢	🟩🟢	🟩🟡	🟩🔴	🟩🔴	🟩🔴	🟩🔴	🟩🟢	🟩🟢	🟩🟢	🟩🔴	🟩🟢	🟥🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢
SC	🟩🟢	🟥🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟩🔴	🟥🔴	🟩🟢	🟩🟡	🟩🟢	🟥🟢	🟩🟢	🟩🔴	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟥🔴	🟩🔴	🟩🟢	🟥🔴
EA	🟩🟢	🟥🟢	🟥🟢	🟥🟢	🟩🔴	🟥🔴	🟥🟢	🟥🟢	🟩🟢	🟩🟢	🟩🟢	🟩🟢	🟥🔴	🟩🔴	🟩🔴	🟩🔴	🟥🟢	🟩🟢	🟩🟢	🟩🟢	🟥🟢	🟩🟢	🟩🔴	🟩🔴

Figure 5: Predictive skill (MSESS, ACC, delta_ACC and CRPSS) and added value of the regional MiKlip ensembles (CCLM_b0 and CCLM_b1) over the eight PRUDENCE regions (cf. Fig. 1) for temperature (left columns), precipitation (middle), and 10m-wind (right) for the multi-annual mean anomalies of lead years 1-5. Red filled boxes indicate negative skill scores, green filled boxes positive skill scores. Green dots denote an added value compared to the global forcing by MPI-ESM-LR, red dots indicate no added value by regionalization, and yellow dots indicate skillscores which are equal in CCLM and MPI-ESM-LR. For MSESS, delta_ACC, and CRPSS we have used the uninitialized historical ensemble as reference dataset.

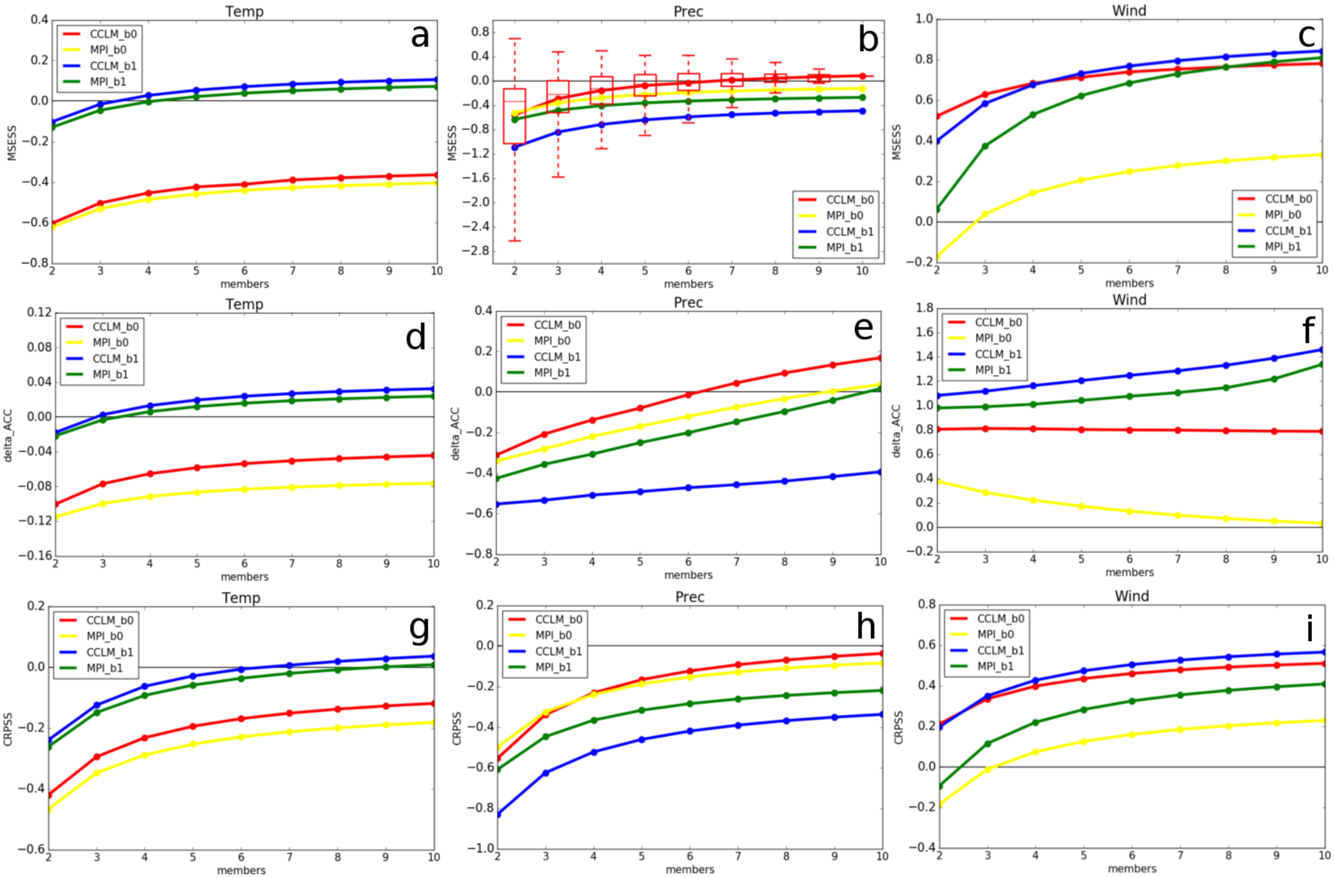


Figure 6: Skill scores for the multi-annual mean anomalies of lead years 1-5 of the CCLM_b0 (red), MPI_b0 (yellow), CCLM_b1 (blue), and MPI_b1 (green) ensembles depending on the ensemble size (x-axis, ranging from 2 to 10 members) over IP (cf. Fig. 1). MSESS for (a) temperature, (b) precipitation, and (c) wind speed; delta_ACC for (d) temperature, (e) precipitation, and (f) wind speed; CRPSS for (g) temperature, (h) precipitation, and (i) wind speed. In (b) box-whisker plots for the skill scores of all n-member combinations are shown. For MSESS, delta_ACC, and CRPSS we have used the uninitialized historical ensemble as reference dataset. Note the different scaling of the y-axis. For details please refer to main text.

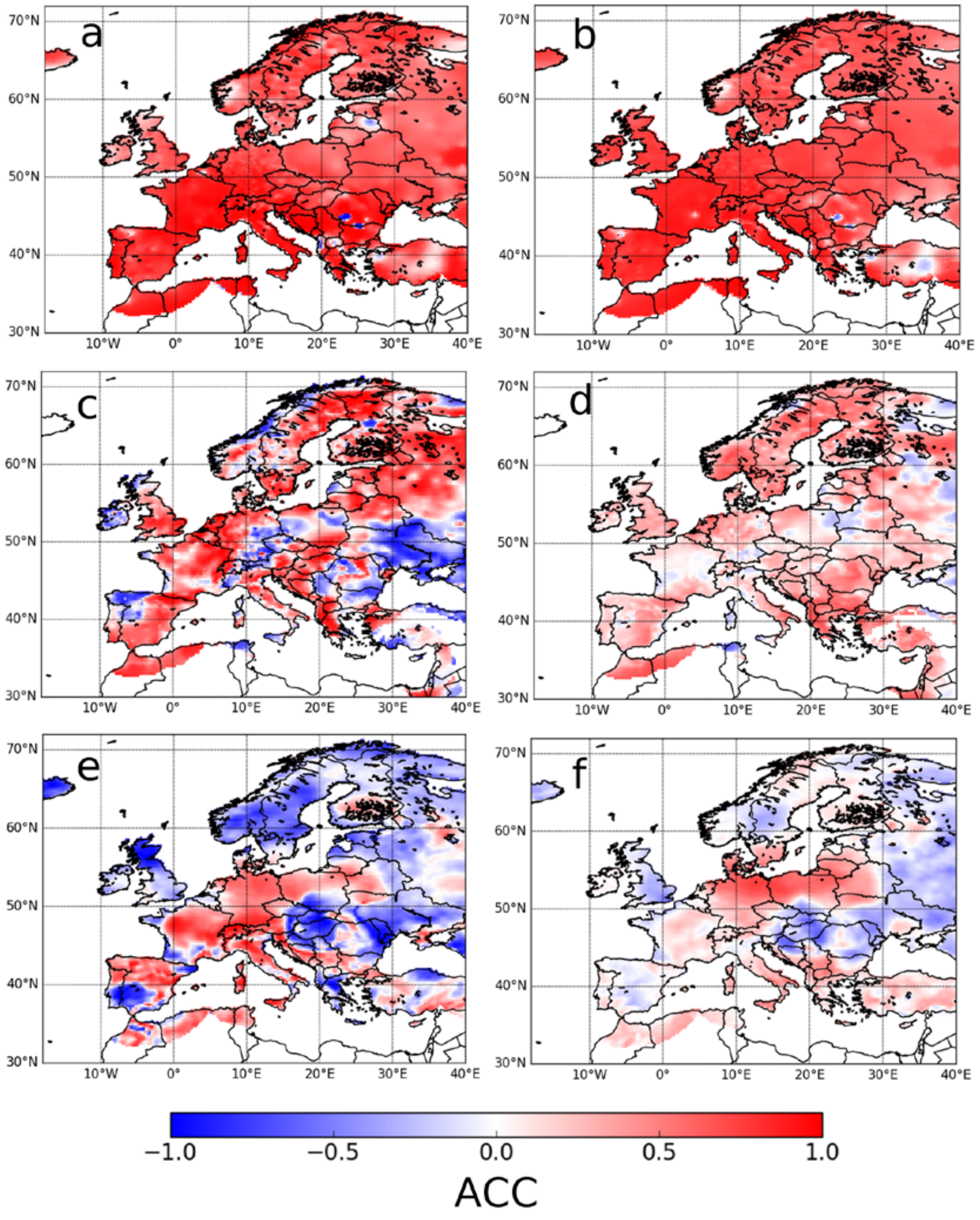


Figure 7: Spatial distribution of the ACC for the multi-annual mean anomalies of lead years 1-5 in MPI_b1 for (a,b) temperature, (c,d) precipitation, and (e,f) wind speed. For the left panels five start years (dec1960, dec1970, dec1980, dec1990, dec2000) have been used, while for the right panels all start years from dec1960 to dec2000 are taken into account. For more detaols see main text.