Point-to-point replies to Reviewer comments for manuscript esd-2017-70

Comment by the authors:

Please find below point-to-point replies to the comments of anonymous Referee #1 and #3. Responses are given in red. Changes in the revised manuscript according to the comments are also given in red. Note that we have replaced old Fig. 2-7 in the revised version according the suggestions of Referee #3. Further, old Fig. 5 is now split into three Tables (Table 1-3 in the revised manuscript). New Figures 2-4 and 6 are attached at the end of this document as Figures R1-R4.

Anonymous Referee #1

In the field of multi annual prediction this paper reflects the state of the art and is the first to demonstrate skill on the regional scale. This manuscript should be published nearly as is. I have only some minor suggested editorial corrections.

Editorial changes:

In section 2 on page 3 the authors should note that anomaly initialization is being used in all ensemble members, global and regional. This is only stated on page 14 and the lack of bias or drift in forecasts is puzzling until then.

Answer: As suggested by the Reviewer we noted that anomaly initialisation is used for all ensemble members in section 2 of the revised manuscript.

In the discussion of the MSE score at the top of page 6, n and N should be defined.

Answer: We thank the Reviewer for this hint. We have clarified this in the revised manuscript.

On page 12 line 20 'chapter 4.1' should be 'section 4.1'

Answer: We have changed „chapter“ to „section“ in the revised version.

Suggest that 'With is respect' be changed to 'With respect to this' on page 13 line 27

Answer: As suggested by the Reviewer, we have reformulated this sentence.

Anonymous Referee #3

General Comment:

The key issues raised by reviewer 3 below concern the quantification of the significance of the results and in particular the quantification of uncertainty. However, we think that these doubts are mainly due to the presentation of the results but not due to the results per se. We have addressed these issues throughout the whole revised manuscript:

- By including significances/uncertainties/quantifications using a bootstrap approach in the revised manuscript (see attached new figures R1-R4, replacing old figures 2, 3, 4 and 7, and specific comments below), we demonstrate that the prediction skill may indeed be significantly improved by dynamical downscaling for the selected meteorological variables in certain regions.
- By including additional quantitative arguments and significances using a t-test

Recommendation to the editor

Suggestions for revision or reasons for rejection (will be published if the paper is accepted for final publication)

The manuscript "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. provides an insight into regional decadal climate prediction activities that have been undertaken within the MiKlip initiative.

The manuscript targets at answering four central research questions:
a) Is there a potential for skillful regional decadal predictions in Europe?
b) Does the regional downscaling provide an added value for decadal predictions?
c) Does the regional decadal predictive skill depend on the ensemble size?
d) How does the sample size affect the skill estimates?

I was not involved in the first round of the review process. I read the manuscript first without having a look at the response letter before to remain unbiased in this respect. I appreciate the effort the authors spent so far to develop the manuscript. However, even though a regional decadal prediction system is something of great interest for the climate science community, I have to suggest rejecting the manuscript given its current shape and scientific focus.
This for a number of general reasons:

1. The authors state that MiKlip is the first effort worldwide establishing a decadal prediction system for the regional scale (I guess they mean the included dynamic downscaling here.). To my knowledge this statement is absolutely correct but this is for a reason. The dynamical downscaling of the decadal predictions is associated with large computational costs. The authors have to prove that these costs are outperformed by benefits in prediction skill for certain variables or phenomena. The manuscript currently fails completely to deliver this evidence:

Answer: We agree with the Reviewer that dynamical downscaling of a large ensemble of decadal hindcasts has high computational costs. However, the aim of this study is to evaluate the dynamical downscaled hindcasts, which is a first step to establish a regional prediction system. Once the regional system is evaluated, only the decadal predictions need to be

downscaled, and not the whole ensemble again. As these predictions will consist of one start year in an operational system, the computational costs are acceptable. There are further justifications for the choice of the dynamical downscaling approach, which are given in our answers to the Reviewer's comments below.

1.1. I do not understand why the authors chose perennial means of temperature, precipitation, and wind as the basis for their validation of a regional decadal prediction system. The variability of such multi-year means of primary meteorological parameters will be mainly determined by low-frequency variability of large-scale circulation patterns (e.g. NAO).

Answer: Multi-year means are used in most studies dealing with the validation of decadal prediction systems (Doblas-Reyes et al., 2013; Marotzke et al., 2016, Meehl et al., 2014 and references therein). Their usage is even recommended by Goddard et al. (2013) and Boer et al. (2016). Therefore, the use of multi-year means in the analysis of decadal predictions is a state-of-the-art approach. The Reviewer is correct in the sense that the low-frequency determines the multi-year variability of the primary meteorological parameters. This is in fact the main source of decadal predictability. The leading mode of variability on this time-scale is the Atlantic Multidecadal Variability (AMO, Smith al al., 2012; Sutton and Hodson, 2005 etc), which has a strong impact on the European climate. We have added a short paragraph in the introduction section of the revised manuscript.

Smith D. M., Scaife A. A., and Kirtman B. P. (2012) What is the current state of scientific knowledge with regard to seasonal and decadal forecasting? Environ. Res. Lett., 7(015602), doi: 10.1088/1748-9326/7/1/015602

There is no need to employ an RCM to model this, this is done (more or less successfully) by the driving GCM. The only benefit provided by the RCM for these variables' means is (hopefully) some correction of the conditional bias (improved variance) due to a better resolution of orography and land-use. However, such (conditional) bias-correction could be done much cheaper by means of statistical downscaling. I would like to read an argumentation why the authors expect regional downscaling to be valuable in this respect.

Answer: We think that dynamical downscaling has many advantages compared to statistical downscaling. First, all variables are physically consistent in dynamical downscaled model runs, which is not the case for statistical downscaling of multiple variables. Second, a statistical method has to be developed for all variables of interest separately, which is associated with high technical efforts. And third, in the statistical downscaling the relationship between large-scale and regional variables is only known for the period for which the method is developed, but it is unclear if it remains the same in future decades. This issue is overcome by the dynamical downscaling due to its physical consistency.
Further, and as mentioned above, the aim of this study is to evaluate the regional system towards a later operationalisation. In future applications, the dynamically downscaled ensemble will be used to analyse user-oriented parameters (e.g. wind gusts, heavy precipitation events potentially causing floods, growth degree days, …), which often consist of a combination of different variables on the regional scale. Hence, ensembles of numerous consistent variables are required for this purpose, which can only be delivered by a dynamical downscaling approach. Finally, the regional decadal hindcasts may be used for impact modelling or as forcing data for hydrological models, which requires physically

consistent data. We have added a brief discussion of the advantages of dynamical downscaling in the introduction section of the revised manuscript.

1.2 Additionally, I have problems interpreting/understanding the "proof" of added value (i.e. skill) the downscaling yields compared to the GCM predictions. I understand that the authors present Fig. 4 & 5 targeting this question. The problem with Fig. 4 is that the key message (RCM prediction better than GCM prediction) is based on qualitative evaluation. I agree that it seems from Fig. 4 that skill scores are shifted up and right in the scatter diagrams when comparing the RCM predictions to the GCM predictions. But this is just a result from eye-balling the plots. There is no direct quantification of the gain (and its significance). In principle it would be possible to show differences between 4a&b, 4c&d, and 4e&f, respectively. However, I do not really suggest to go for this solution (calculating the differences of these scatter diagrams). I would rather propose to present skill maps (as done for Fig 2 & 3) again, that is a map plot showing skill of RCM-predictions with the GCM-predictions as reference forecast.
Answer: We thank the Reviewer for this comment, which includes helpful suggestions. Following the Reviewer's suggestion, we have replaced the old Fig. 4 by a map plot showing the MSESS of the RCM-predictions using the MPI predictions as reference dataset. This enabled us to quantify where the dynamical downscaling significantly improves the prediction skill. We have changed the text accordingly in section 4.2 of the revised manuscript.

Fig.5 is problematic, too. First, I do not understand what the authors indicate by the background color. They say it is the "absolute skill" (page 18 line 32). Is this supposed to mean that the background color indicates whether there is skill over the climatological forecast, while the dots mark skill over the GCM-predictions? If so, that is fine but should be explicitely written in the manuscript (and the figure caption). And I do not understand the difference between ACC and delta_ACCin this diagram. An additional problem of this plot (as with most of the analyses) is that it does not take the uncertainty into account (see my dedicated issue below). When do the authors consider skill scores to be equal??? Strictly speaking, every prediction should be considered equal that is not proven to be significantly better than the reference forecast. My assumption is, that this would lead to a totally yellow table diagram in Fig. 5.
Answer: We agree that Fig. 5 is a very busy graphic. For the revised manuscript, we split the information into three tables, including quantitative numbers and use the t-test to derive a measure where the skill differences are significant. Additionally, we decided to remove delta_ACC in this section and include results from the Murphy decomposition instead (see also last comment).

My general doubt regarding the benefit from dynamical downscaling is further supported by Fig. 6. Even though compiled for another research question (the ensemble size-dependent bias), this figure shows impressively (for the example of the IP-region) that the downscaling hardly yields any benefit. If the authors would quantify the uncertainties, I am pretty sure that all RCM-curves would be situated within the confidence intervals of the respective GCM curves.
Answer: The box-whisker plots shown in old Fig. 6b represent the ensemble spread resulting from the choice of the ensemble members. As this spread represents a part of the uncertainty, we agree with the Reviewer that no significant differentiation of the skill scores can be determined for small ensemble sizes. However, for an ensemble size of seven and more, the

confidence interval strongly decreases. We have now added box-whiskers in new Fig. 5d-5f in the revised manuscript for different purposes (see revised main text). In particular from new Fig. 5f it is visible that for the ensembles of more than seven members the CCLM_b0 curve is well above the confidence interval of MPI_b0. Hence, as already shown in the new Fig. 4 (see also answer to comment 1.2 above), the prediction skill may indeed be significantly be improved when downscaling the full ensemble of ten members depending on the variable and the region.

All this sums up to my point of view that the manuscript fails to provide evidence of (significantly) added value achieved by downscaling for the analyzed variables. From my point of view, the great potential of employing a non-hydrostatic RCM for decadal predictions could be in the representation of extremes, especially for precipitation. However, frequencies of extremes or something similar are not addressed by the manuscript.
Answer: Regarding the Reviewer's criticism given in comment 1.1 and 1.2, we can understand the doubts of the Reviewer regarding the added value of downscaling. However, we think that these doubts were mainly due to presentation of the results, and not due to the results per se. By including significances/uncertainties/quantifications in the revised manuscript (see our answers above), we could demonstrate that the prediction skill may indeed be significantly improved by dynamical downscaling.
Again, the aim of our study is to evaluate the regional ensemble. Extremes and other user-relevant parameters will be examined in future applications.

2. A point already issued by one of the last round's reviewers (and explicitely mentioned in the editor's decision) is the low number (5) of initializations. The authors present figure 7 in the revised manuscript and state that these plots prove a general compliance of the results derived from five initializations with those based on 41 initializations. This is true for temperature where skill basically exists for whole Europe due to extrenally forced global warming. However, for precipitation and wind there are some important differences between results based on only 5 and 41 initializations, respectively. I resist from naming them explicitely, the key message here is that the uncertainty associated with all skill scores calculated for the manuscript is huge. And of course the skill estimates based on 41 initializations in Fig. 7 still feature substantial uncertainty, they are not the ultimate "true" skill. This is the essential problem of the manuscript: uncertainty is never quantified. Without such quantification of uncertainty/significance, the results cannot be taken scientifically serious.
Hence, the manuscript is not able to achieve its central aim that is providing evidence or at least significant supportive indication of a potential benefit from dynamical downscaling for decadal climate prediction.
Answer: The Reviewer is correct that a low number of hindcast starting dates poses a serious problem in the analysis of decadal predictions in general (and in this study in particular). This was recognized after CMIP5 (Meehl et al., 2013) and changed for CMIP6 (Boer et al., 2016). As already stated in the conclusions of the paper, a result of the efforts presented in this manuscript is to use annual starting dates on the regional scale for further hindcast generations. New Figure 6 and research question d) explicitly address this issue. We think, that this part is important for the paper (and was requested by the other reviewers), i.e, to show how far the results obtained for the other research questions can be transferred to an ensemble with a larger sample size. To address the concern raised by the Reviewer regarding the estimation of the uncertainty, we have now included the significance at the 95%-level of the resulting skill scores derived from a bootstrapping approach (see Fig. R4 and others). The results provide evidence of the statements made in our paper, that there are regions which

provide a significant skill. Fig. R4 is included as new Fig. 6 in the revised manuscript and we have adapted the text accordingly.

3. The scientific questions c) and d), both targeting the ensemble-size-dependent bias of (regional) prediction skill are not worth tackling in a way done by this manuscript. From my point of view, the manuscript does not add anything scientifically new to the literature in this respect. Of course, it is absolutely important to account for this skill-bias when comparing predictions made by ensembles of differing size and it might be helpful to indicate this bias even for analyses of ensembles with same size in order to quantify potential skill of (hypothically) larger ensembles. But there is no doubt about the general question if there is a bias: there is! The authors should have a look into original papers addressing this issue, such as Murphy (1990), Ferro (2007), and Sienz et al. (2016; there might by older papers than Sienz et al., tackling the bias of MSE/RMSE but I currently remember only this rather recent one). They will learn that the skill bias for all metrics used is essentially related to the variance of the target parameter and the actual bias is a result from the additional noise component due to an inaccurate estimate of the signal due to the limited ensemble size. The theoretical behaviour of the bias depending on ensemble-size is thus also known and can be estimated with existing approaches given by these papers. By the way, this disproves the author's statement that only the CRPSS could be corrected, which is why they refrained from using a de-biased version of the CRPSS in order to maintain comparability to the other metrics (page 20 lines 26-30).

Murphy, J.M. (1990): Assessment of the practical utility of extended range ensemble forecasts. Q.J.R. Meteorol. Soc., 116: 89-125. doi:10.1002/qj.49711649105
Ferro, C.A.T. (2007): Comparing Probabilistic forecasting systems with the brier score. Wea. Forecasting, 22: 1076-1088. doi:10.1175/WAF1034.1
Sienz, F., Müller, W.A., Pohlmann, H (2016): Ensemble size impact on the decadal predictive skill assessment. Meteorol. Z., 25(6): 645-655. doi:10.1127/metz/2016/0670

Answer: From our point of view, the ensemble size (e.g. number of realizations) and the sample size (which takes the number of starting dates and respective climate conditions into account) pose different research questions. This is in agreement with the findings by Sienz et al. (2016). The previous studies mentioned the use of either synthetic data (Sienz et al., 2016) or data from weather prediction. In this manuscript, the intention is to show the effect of increasing the ensemble size in actual (regional and global) decadal predictions. It was not the intention here to estimate the limit for an infinite ensemble size. The current decadal prediction systems rarely use large ensemble sizes. For instance, the CMIP5 recommendations proposed an ensemble size of three members for each starting year (Taylor et al., 2012). In new Figure 5, we show that such a low ensemble size of three members is clearly insufficient. Furthermore, the results indicate that the minimum required number of members depends both on the meteorological parameter and on the chosen metric. We agree that the variability between the sub-ensembles for very small numbers of members makes the results indistinguishable – as shown in new figures 5d-f. But we could also provide evidence that with a higher number of members the results for some parameters and metric become distinguishable. We have included box-whisker plots for additional ensembles and variables in the new Figures 5d-f, in order to allow a better quantification of the needed minimum ensemble size. In the revised version of the manuscript, we have improved the explanation.

4. The authors present their analyses based on three metrics, that is the MSESS, the ACC, and the CRPSS. However, there is hardly any information on why the authors chose these three metrics: What is the additional information that is provided by ACC and CRPSS

compared to using the MSESS only? What do we learn from differences between the various metrcis presented in Fig. 5 (e.g. What does it mean that CRPSS for the FR-region wrt wind is positive for the RCM-prediction compared to the GCM-prediction but negative for all other skill metrics?). There are a few hints in Sec. 3.2 when introducing the metrics, however, some of them are rather misleading. This is especially true for the CRPSS which does not only contain information about the "reliability" of forecasts but also about "resolution" (see page 15 line 16). Additionally, it does contain much more information than only checking whether the ensemble spread is an adequate representation of forecast uncertainty (compare page 20 lines 25-26). Last but not least, the authors do not consider the relationship between the different metrics. This is most relevent for the MSE(SS) and the ACC. Following Murphy (1988) the MSE for a forecast f measured against the obervation o can be decomposed into:

MSE(f,o)=(<f> - <o>)^2 + s_f^2 + s_o^2 - 2 s_fs_o R_{fo}

where <f> and <o> are the means of f and o, respectively, s_f^2 and s_o^2 are the sample variance of f and o, respectively, and R_{fo} is the sample correlation of f and o. As soon as anomalies are considered, as done in the study by Reyers et al., the first term becomes 0 for all instances. A positive MSESS (meaning a lower MSE than the reference forecast) hence can be achieved by a smaller forecast variance (this is where the ensemble size is relevant, by the way => larger ensembles yield smaller noise variance) or an increase in correlation. The latter is additionally calculated for the present study, although - as just described - closely linked to the MSE(SS). This might make sense if drawing the right conclusions, especially for cases where ACC and MSESS develop differently. However, currently it seems that the study just somehow "quantifies skill" (whatever that means) without any attempt to assess what the reasons are why the RCM is performing better/worse than the GCM predictions (and still failing because not quantifying the uncertainty). I suggest that the authors present less skill metrics (maximum of two) but clearly think about about which to choose and really spend some effort to disentangle the different meanings provided by these. One natural choice would be to choose MSESS and CRPSS which comprises a maximum of information. However, it might be better to stay with deterministic scores only. This would be the MSE(SS) as primary verification metric with the ACC as complementary metric in order to assess whether it is really the signal that is predicted better/worse or whether it is only a change in forecast variance (which might be the case when comparing RCM and GCM predictions).

Murphy, A.H. (1988): Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient. Mon. Wea. Rev., 116, 2417-2425, doi:10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2

Answer: We thank the Reviewer for this helpful comment. Regarding the choice of metrics and their application, the authors follow the recommendations for the verification of decadal predictions as proposed by Goddard et al. (2013) as cited in the manuscript (page 5, lines 15f). We agree with the Reviewer that the differences and relations of the metrics should be explained in more detail. Following his/her suggestions, we have omitted the CRPSS in the revised version of the manuscript and focus only on MSESS and ACC. Further, we now interpret our results based on the Murphy decomposition and analyse in more detail the contributions of the correlation and the conditional bias to the MSESS skill. In addition, we have clarified where the downscaling is able to improve the skill, which will certainly improve the scientific validity of our study (see also answers to comments above). The Murphy decomposition is described in chapter 3.2 of the revised version.

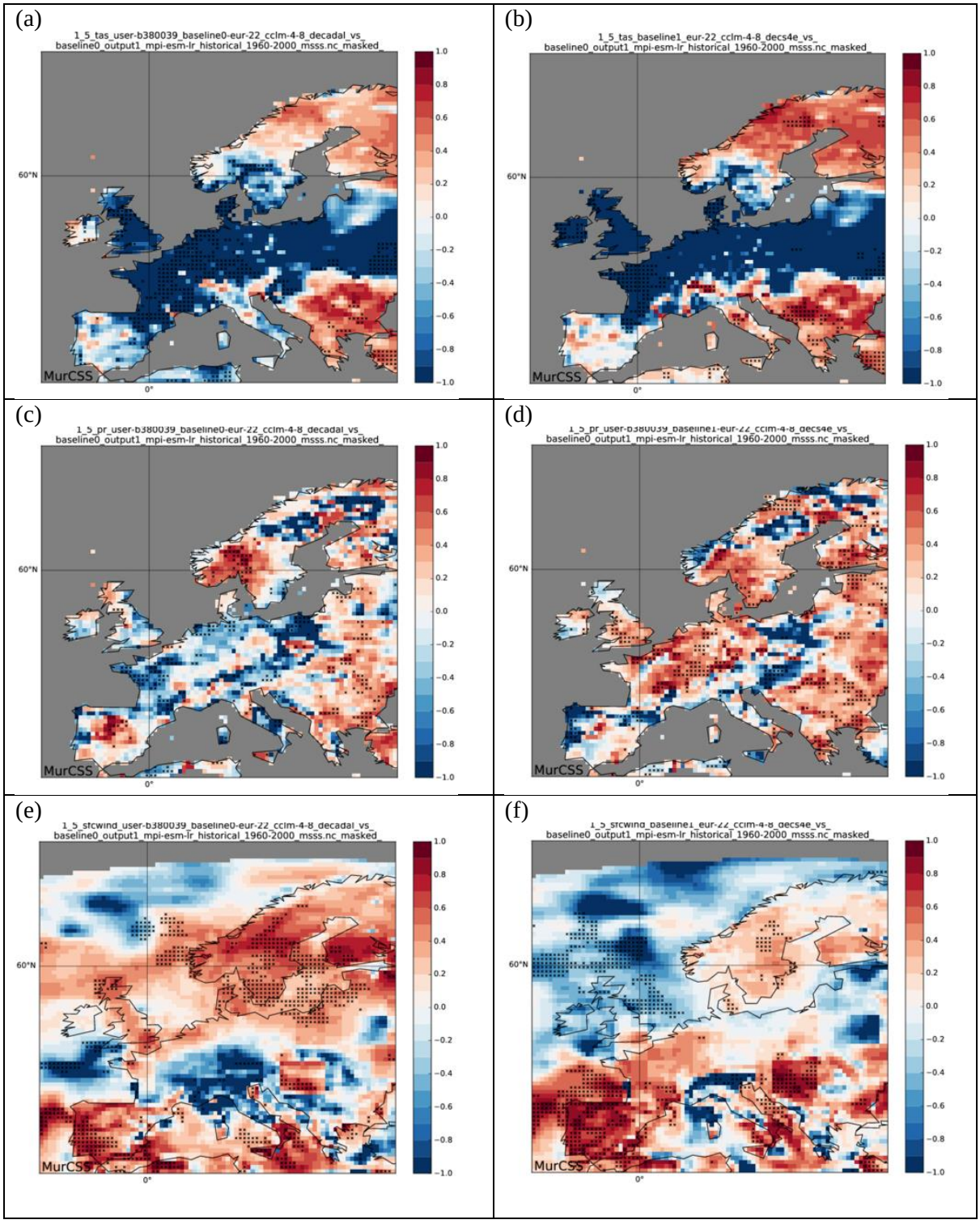

Figure R1: Spatial distribution of the MSESS for the multi-annual mean anomalies of lead years 1-5 for (a) temperature in CCLM_b0, (b) temperature in CCLM_b1, (c) precipitation in CCLM_b0, (d) precipitation in CCLM_b1, (e) wind speed in CCLM_b0, and (f) wind speed in CCLM_b1. As reference dataset, we have used the uninitialized historical ensemble. The black dots indicate significant skill at the 95% level.

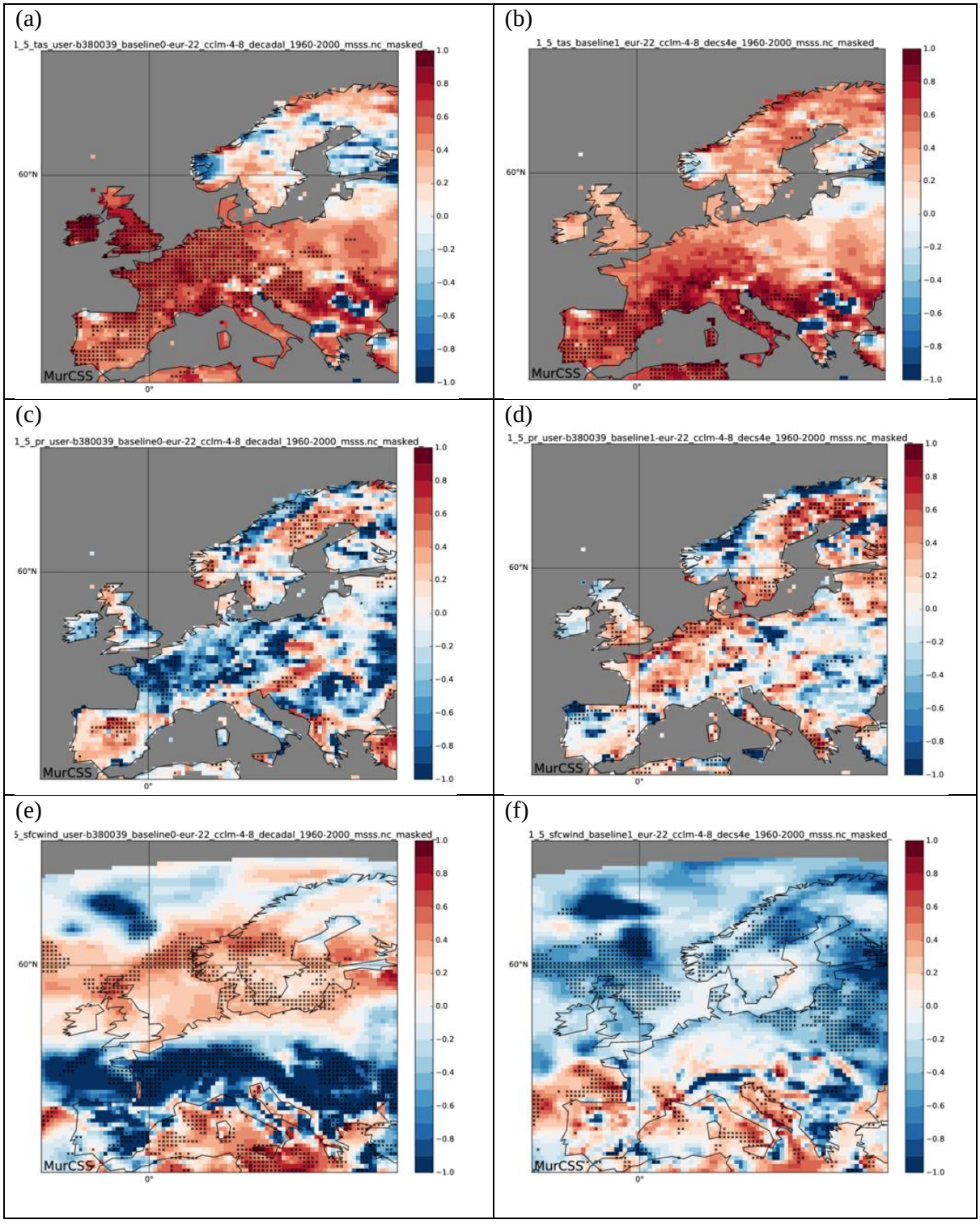

Figure R2: Spatial distribution of the MSESS for the multi-annual mean anomalies of lead years 1-5 for (a) temperature in CCLM_b0, (b) temperature in CCLM_b1, (c) precipitation in CCLM_b0, (d) precipitation in CCLM_b1, (e) wind speed in CCLM_b0, and (f) wind speed in CCLM_b1. As reference dataset, we have used the climatology. The black dots indicate significant skill at the 95% level.

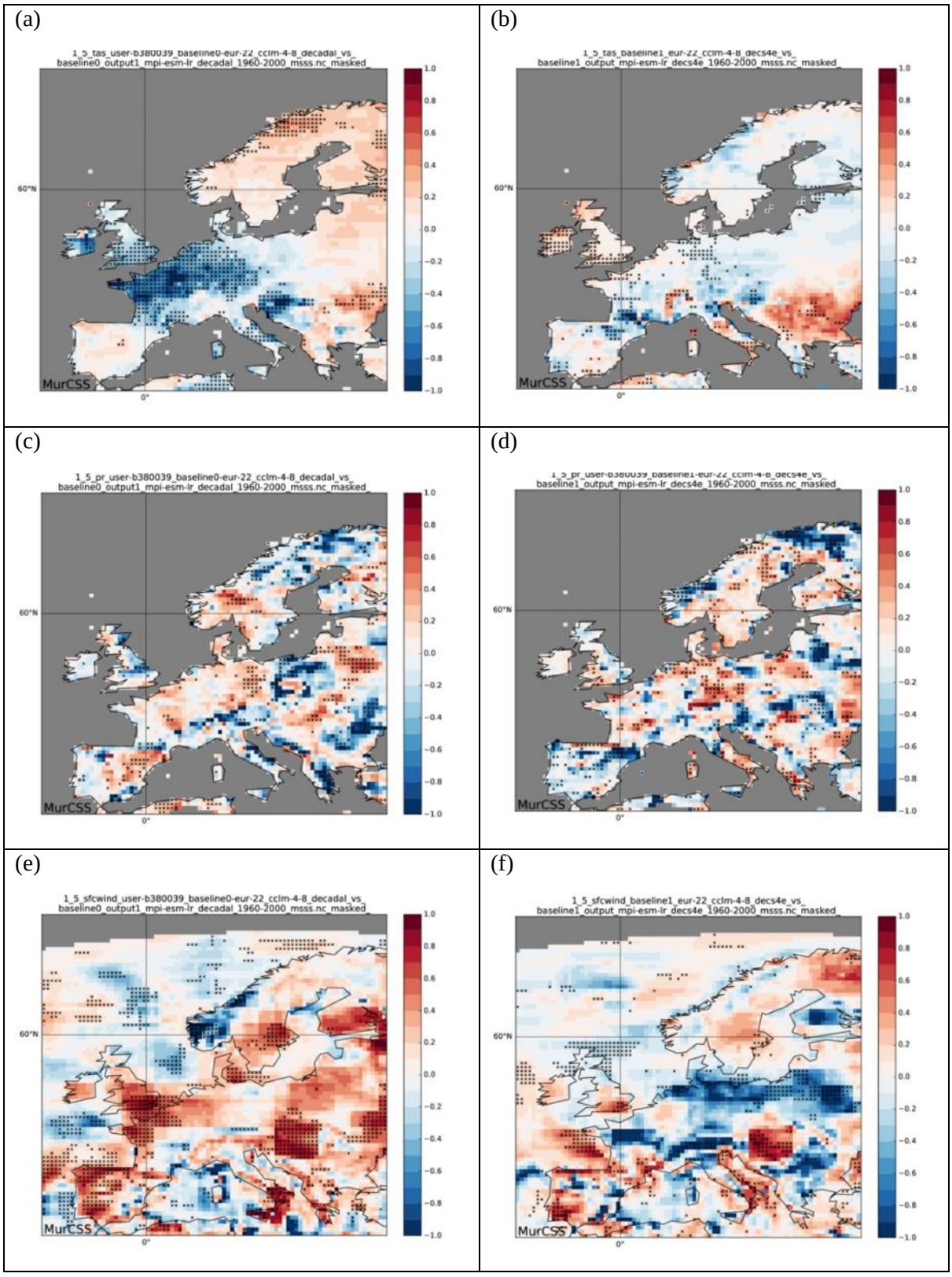

Figure R3: Spatial distribution of the MSESS for the multi-annual mean anomalies of lead years 1-5 for (a) temperature in CCLM_b0, (b) temperature in CCLM_b1, (c) precipitation in CCLM_b0, (d) precipitation in CCLM_b1, (e) wind speed in CCLM_b0, and (f) wind speed in CCLM_b1. As reference dataset, we have used the respective global MPI data. The black dots indicate significant skill at the 95% level.

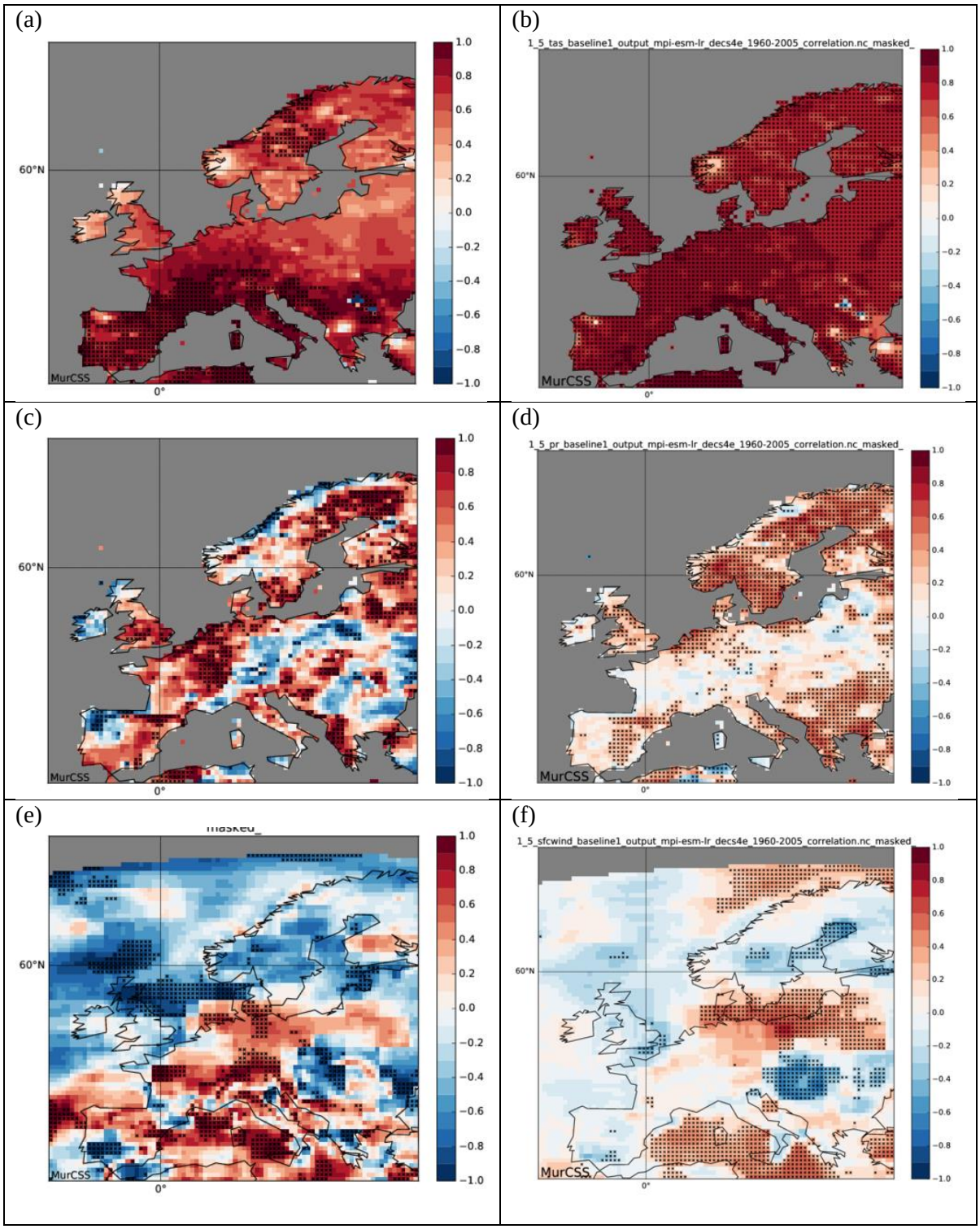

Figure R4: Spatial distribution of the ACC for the multi-annual mean anomalies of lead years 1-5 in MPI_b1 for (a, b) temperature, (c, d) precipitation, and (e, f) wind speed. For the left panels, five start years (dec1960, dec1970, dec1980, dec1990, and dec2000) have been used, while for the right panels all start years from dec1960 to dec2000 are taken into account. The black dots indicate significant skill at the 95% level.

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. 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 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 sub-polar gyre (Matei et al., 2012; Yeager et al., 2012; Robson et al., 2013). Recent studies suggest that in particular the Atlantic multi-decadal variability, which is strongly linked to the AMOC, is a major source of decadal predictability (Smith et al., 2012; Pohlmann et al., 2013a). As such low-frequent variability patterns may affect the climate globally, perennial means of meteorological parameters might be predictable several years ahead. Numerous 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. (2015) 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, and an outstanding large ensemble was regionalised by dynamical downscaling with regional climate models. Although being computational expensive, dynamical downscaling has

many advantages compared to other downscaling methods. For example, all output variables are physically consistent in dynamically downscaled model runs, which is particularly important when using decadal predictions for impact modelling, hydrological simulations, or user-oriented parameters. 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 aspread 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., 2009; 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. Hence, the downscaled hindcasts also inherit the applied anomaly initialization of the global ensembles. 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 metrics are used to evaluate the performance of the global and regional hindcast ensembles and to address the four key questions: the mean squared error skill score (MSESS) and the anomaly correlation coefficient (ACC), which are both measures for the forecast accuracy. 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 deterministic MSESS (Goddard, 2013) is defined as

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

with

$$MSE = \frac{1}{N}\sum_{i}^{N}(\bar{X}_i - O_i)^2$$

where i=1, …, N is the time index, MSE_{hind} is the mean squared error (MSE) between the ensemble mean of the dynamical downscaled hindcasts (X_i) and the verification data, and MSE_{ref} is the mean squared error of a reference dataset. In this study, either the uninitialised historical simulations, the climatology, or the global initialised hindcasts are used as reference. A positive MSESS means that the hindcasts are closer to the verification dataset than the reference, indicating that the initialisation and downscaling lead to higher accuracy in predicting observed values.

Following Murphy (1988) and given that anomalies are used (as in this study), the MSESS with the climatology as reference can be decomposed as follows:

$$MSESS = r_{X,O}^2 - \left| r_{X,O} - \frac{S_X}{S_O} \right|^2$$

Where $r_{X,O}$ is equivalent to the ACC described below and the sample variances of the simulation ensembles (S_X) and the observations (S_O), respectively. The second term is the conditional bias.

$$CB = \left| r_{XO} - \frac{S_X}{S_O} \right|$$

Hence, MSESS is smaller than the correlation in case there is a conditional bias, for which the optimal value is 0. CB depends on the ratio of the standard deviation between the ensemble and the observation. The improvement or added value of CB is calculated according to Kadow et al. (2015) as:

$$CB_{AV} = \left| r_{ref,obs} - \left(\frac{S_{ref}}{S_O} \right) \right| - \left| r_{hind,obs} - \left(\frac{S_{hind}}{S_O} \right) \right|$$

The correlation or ACC (for anomaly correlation coefficient; 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).

The significance of the skill scores is determined by using a bootstrapping approach at the 95% level (Kadow et al., 2015) and a t-test of the respective distributions.

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, and at some grid points this prediction skill is significant. 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 and partly significant 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 and significant 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. It is significant for most regions in Western Europe in CCLM_b0 and large parts of Southern Europe in CCLM_b1 (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. Here, skill scores are often significant. 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).

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. With this aim, we use MPI_b0 and MPI_b1 as reference datasets for the MSESS shown in Fig. 4 (see also section 3.2).

Generally, significant improvements in the prediction skill by dynamical downscaling are restricted to limited geographical areas, and they strongly depend on the variable. For temperature a significant added value of downscaling is found over Scandinavia and Southeast Europe in CCLM_b0, and over Southeast Europe and the British Isles in CCLM_b1 (Fig. 4a and 4b). Therefore, the regionalization typically provides an improvement in regions where the global hindcasts show mostly medium to lower skill. In some regions with already high skill of the MPI-ESM hindcasts there is no improvement by the downscaling. This includes, for example, an area from the British Isles over France to Germany in baseline0 and regions along the Western Mediterranean coast in baseline1. However, the global model outperforms the regional model over large parts of Northwest Europe in the baseline0 ensemble (see also Table 1).

Again, rather patchy MSESS fields are obtained for precipitation. Nevertheless, there are several regions with significantly improved prediction skills in the CCLM ensembles compared to MPI_b0 and MPI_b1 (Fig. 4c and 4d). For example, both CCLM_b0 and CCLM_b1 reveal significant positive MSESS over Eastern Germany and over parts of Scandinavia.

The added value of downscaling is most pronounced for wind in CCLM_b0 (Fig. 4e). Significant improvements of the MSESS are detected for Southeast Europe, Italy, Scandinavia, and for coastal areas of Iberia, France and the United Kingdom. Areas with an added value of downscaling are also existent in CCLM_b1 (Fig. 4f), but visibly reduced compared to CCLM_b0.

In the Tables 1 – 3 we summarize the analysis of skill (cf. Fig. 3) and added value (cf. Fig. 4) of the regional hindcasts compared to the climatology for the three variables as spatial means over the PRUDENCE regions (cf. Fig. 1). The tables display the MSESS as well as its components correlation (ACC) and conditional bias (CB), according to the Murphy decomposition (cf. chapter 3.2). The colouring of the cells in the left half of the tables display positive (green), negative (red) and neutral skill (white) derived from the results shown in Fig. 3 for the MSESS. The thresholds for the colouring are above/below +/-0.25 for the MSESS, and +/-0.4 for ACC. Since the optimal value for the conditional bias is zero, a low CB is depicted for the absolute |CB| being below 0.2 (green), and a high CB for |CB| being larger than 0.3 (red). The colouring for the added value in the right

half of the tables corresponds to the sign of the skill difference (dynamical downscaled minus global MPI; green: positive; red: negative) in a region in connection with a t-Test, to determine the significance of the changes in the distribution at the 80%-level.

For temperature (Table 1), the MSESS and ACC are high in all regions except Scandinavia (SC). The correlation is above 0.8 in CCLM_b0 for the north-western part of Europe – namely the British Isles (BI), France (FR), Mid-Europe (ME) and the Alps (AL). For CCLM_b1 the highest correlation is found more in the southern part of the region, namely in France (FR), the Iberian Peninsula (IP), the Alps (AL) and the Mediterranean (MD). The CB is low for most areas, except France (FR) and Scandinavia (SC) in CCLM_b0. The MSESS is high due to the high ACC and the low CB. MSESS and ACC are higher and CB lower for b1 than for b0 over whole Europe (EU). A significant added value of the regionalization of basline0 occurs for Scandinavia (SC) and Eastern Europe (EA) as well as for Europe (EU). In these regions CB is significantly lower for CCLM compared to the MPI-ESM hindcasts. CCLM_b1 shows an added value for the regions British Isles (BI), the Mediterranean (MD) and Eastern Europe (EA) as well as over the entire domain (EU). The main cause is again a reduced conditional bias.

The MSESS for precipitation (Table 2) is much lower than for temperature. It is mostly negative for CCLM_b0 and only slightly positive for some regions in CCLM_b1. For CCLM_b0 a significant positive correlation is only found for the Iberian Peninsula (IP). The CB is generally negative (except for IP), which is inherited from the global hindcasts, since for MPI_b0 CB is even lower than for CCLM_b0. Over the whole domain the added value is low and not significant. For CCLM_b1 a positive ACC is uncovered for France /FR), Mid-Europe (ME) and Scandinavia (SC). The conditional bias is much lower than for CCLM_b0 and only slightly negative in most areas. An added value in relation to MPI_b1 occurs in those regions, where ACC of CCLM_b1 is positive. There is also an added value for the whole domain (EU), but it is only significant for ACC.

The highest ACC for the 10m-wind speed is revealed for northern Europe (BI and SC) as well as in the Mediterranean (MD) for CCLM_b0, and in the southern regions for CCLM_b1 (Table 3). These are all regions with low to moderate conditional bias. CB is strongly negative in regions with a low skill. Interestingly, the added value for the MSESS of CCLM_b0 is significant in all regions, and for most regions a significant improvement is also visible with respect to ACC and CB. For CCLM_b1, on the other hand, significant improvements of MSESS and ACC compared to MPI_b1 are only found for a few regions.

We conclude that regional downscaling indeed may provide an added value for decadal predictions over Europe, both on individual grid points as well as for spatial means. However, this added value is not systematic but depends on variable and region.

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. (2015) 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 temperatures 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. Fig. 5 exhibits the dependency of MSESS and delta_ACC when compared to the historical simulations 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).

Enhanced predictive skill can be observed when the number of members is stepwise increased for both the global and the regional hindcast ensembles. MSESS shows a rather logarithmic relationship with increasing *n*, depicting the highest skill scores for the 10 member ensembles for all three variables (Figure 5a-c). On the other hand, the lowest skill scores are always found for the 2-member ensembles. This ensemble size dependency of MSESS 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 values to positive values in one or more of the ensembles (e.g. Fig. 5a and 5c). 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. 5d-f). Nevertheless, there are also examples for delta_ACC where the ensemble size dependency is similar to that of MSESS, like e.g. for temperature (Fig. 5d). These results suggest that a decadal prediction system generally benefits from larger ensemble sizes, either in terms of more skilful 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 both 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. 5c and 5f). This is mainly due to higher skill scores over orographic structured terrains of IP in CLM_b0 compared to MPI_b0 (cf. Fig. 4).

For ensembles with less than 10 members, the skill scores of all possible *n*-member ensemble combinations are averaged. For selected ensembles, box-whisker plots of these n-member combinations are shown for delta_ACC in Fig. 5d-f. 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, delta_ACC for wind in CCLM_b0

(MPI_b0) varies between 0 and +1.6 (-0.1 and +1.1) for the 2-member ensembles (Fig. 5f). Even for the 7-member ensemble, results can differ quite strongly depending on the selection of the ensemble members. Similar results are found for temperature and precipitation. These findings clearly demonstrate the necessity of using large ensembles to reduce uncertainties. Further, only for high numbers of ensemble members (eight or more), the delta_ACC curve for CCLM_b0 is above the range of uncertainty of MPI_b0 in case of precipitation (Fig. 5e) and wind (Fig. 5f). This indicates that the prediction skill may only be significantly improved when the whole ensemble is dynamically downscaled. The same applies for the improvement from baseline0 to baseline1 in case of temperature (Fig. 5d).

We conclude that the predictive skill 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 in all hindcasts (see Fig. 5a-c). 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. 6 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 and larger areas with significant skill scores. 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 northeastern 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 and significance of the results. 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 and to some extent the significances of the regional distribution. Therefore, the analysis of the larger

sample size confirms the qualitative findings from section 4.1. 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. Two skill scores are used to quantify the accuracy 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 to some extend significance of skill 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. However, 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 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 and significance 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 respect to this, 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. (2015), 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 (Khodayar 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, and Christopher Kadow and Sebastian Illing for providing the MiKlip Central Evaluation System (MiKlip CES). 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, 2009.

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, 2015.

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.

Murphy, A. H.: Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient, Mon. Wea. Rev., 116, 2417–2424, doi:10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2, 1988.

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.

Smith, D. M., Scaife, A. A., and Kirtman, B. P.: What is the current state of scientific knowledge with regard to seasonal and decadal forecasting? Environ. Res. Lett., 5, 015602, doi:10.1088/1748-9326/7/1/015602, 2012.

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.

Tables

Table 1: MSESS, ACC and conditional bias for the CCLM ensembles (left half) and added value compared to the global MPI ensembles (right half) for b0 (upper half) and b1 (lower half) for temperature averaged over the eight Prudence regions (cf. Fig. 1) and whole Europe (EU). In the left half green (red) colouring display MSESS values above +0.3 (below -0.3), ACC values above +0.4 (below -0.4), and CB values above -0.25 and below +0.25 (lower than -0.25 or higher than +0.25). In the right half green colouring corresponds to a significant improvement (deterioration) by dynamical downscaling using a t-test.

temp	Skill			Added Value		
b0	MSESS	ACC	CB	MSESS	ACC	CB
BI	0.75	0.90	0.15	-0.04	-0.02	-0.02
FR	0.66	0.89	0.34	-0.12	-0.05	-0.04
ME	0.56	0.80	0.23	-0.11	-0.05	-0.05
AL	0.57	0.82	0.10	-0.05	-0.04	0.02
IP	0.50	0.76	0.22	0.01	0.03	-0.08
MD	0.42	0.75	-0.04	0.01	-0.02	0.05
EA	0.45	0.72	-0.16	0.05	0.00	0.11
SC	0.02	0.45	-0.36	0.20	0.07	0.16
EU	0.35	0.67	-0.09	0.05	0.01	0.09
temp	Skill			Added Value		
b1	MSESS	ACC	CB	MSESS	ACC	CB
BI	0.34	0.64	0.25	0.12	0.15	-0.07
FR	0.63	0.87	0.34	-0.03	-0.01	-0.01
ME	0.51	0.75	0.19	-0.03	-0.02	-0.02
AL	0.68	0.89	0.04	-0.03	-0.01	0.06
IP	0.62	0.83	0.20	-0.01	0.01	-0.03
MD	0.52	0.82	-0.09	0.08	-0.01	0.07
EA	0.37	0.66	-0.20	0.05	-0.04	0.15
SC	0.24	0.56	-0.18	-0.03	-0.03	-0.09
EU	0.40	0.69	-0.05	0.01	-0.01	0.01

Table 2: As Table 1, but for precipitation.

prec	Skill			Added Value		
b0	MSESS	ACC	CB	MSESS	ACC	CB
BI	-0.17	-0.04	-0.39	-0.09	-0.09	-0.13
FR	-0.64	-0.55	-0.87	0.07	0.09	0.01
ME	-0.49	-0.31	-0.72	0.18	0.26	0.18
AL	-0.27	-0.10	-0.46	-0.14	-0.04	-0.19
IP	0.12	0.41	0.06	0.05	0.17	0.12
MD	-0.08	0.20	-0.25	-0.13	-0.07	-0.16
EA	-0.28	0.02	-0.45	-0.05	-0.09	-0.05
SC	-0.06	0.16	-0.27	0.02	-0.01	0.05
EU	-0.17	0.06	-0.35	0.00	0.01	0.01
prec	Skill			Added Value		
b1	MSESS	ACC	CB	MSESS	ACC	CB
BI	0.07	0.21	-0.04	0.01	0.01	-0.01
FR	0.21	0.51	-0.04	0.04	0.05	-0.02
ME	0.20	0.54	0.01	0.16	0.20	0.01
AL	-0.01	0.19	-0.11	-0.06	-0.05	-0.11
IP	-0.09	0.12	-0.20	-0.15	-0.13	-0.16
MD	0.05	0.29	-0.13	0.03	0.03	-0.02
EA	-0.11	0.05	-0.32	0.05	-0.01	0.03
SC	0.06	0.41	-0.17	0.00	0.05	0.04
EU	0.03	0.28	-0.16	0.02	0.02	0.00

Table 3; As Table 1, but for wind speed.

wind	Skill			Added Value		
b0	MSESS	ACC	CB	MSESS	ACC	CB
BI	0.21	0.50	0.19	0.23	0.09	-0.12
FR	-0.77	-0.63	-0.88	0.09	-0.04	0.06
ME	-0.22	-0.04	-0.34	0.23	0.13	0.23
AL	-0.69	-0.45	-0.81	0.07	0.06	0.03
IP	-0.36	-0.08	-0.52	0.26	0.17	0.18
MD	0.06	0.38	-0.10	0.12	0.02	0.18
EA	-0.46	-0.21	-0.59	0.36	0.10	0.31
SC	0.19	0.50	0.17	0.16	0.09	0.02
EU	-0.11	0.16	-0.20	0.15	0.07	0.21
wind	skill			Added Value		
b1	MSESS	ACC	CB	MSESS	ACC	CB
BI	-0.35	-0.57	-0.80	0.02	-0.09	-0.01
FR	-0.01	0.28	-0.22	0.00	0.07	-0.05
ME	-0.12	-0.06	-0.38	-0.24	-0.41	-0.29
AL	-0.14	0.32	-0.30	-0.30	-0.13	-0.22
IP	0.05	0.29	-0.17	0.19	0.21	0.18
MD	0.07	0.41	-0.06	0.02	0.05	-0.03
EA	-0.38	-0.23	-0.69	-0.09	-0.07	-0.21
SC	-0.37	-0.34	-0.67	0.06	-0.05	0.02
EU	-0.23	-0.11	-0.50	0.00	-0.02	-0.04

Figures

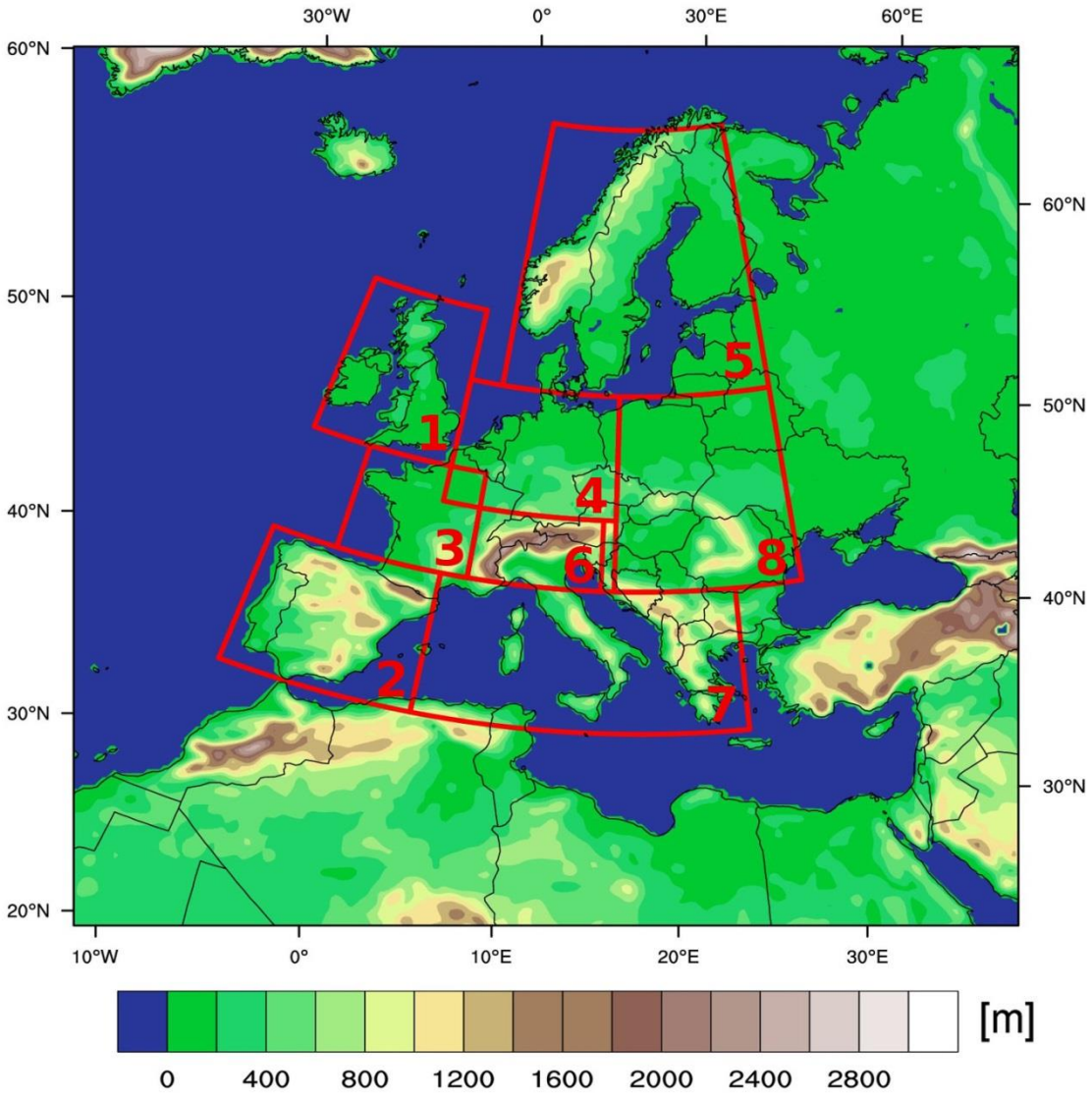

Figure 1: CCLM modelling domain (= EURO-CORDEX domain): Modell orography and PRUDENCE regions. 1: British Isles BI; 2: Iberian Peninsula IP; 3: France FR; 4: Mid-Europe ME; 5: Scandinavia SC; 6: Alps AL; 7: Mediterranean MD; 8: Eastern Europe EA.

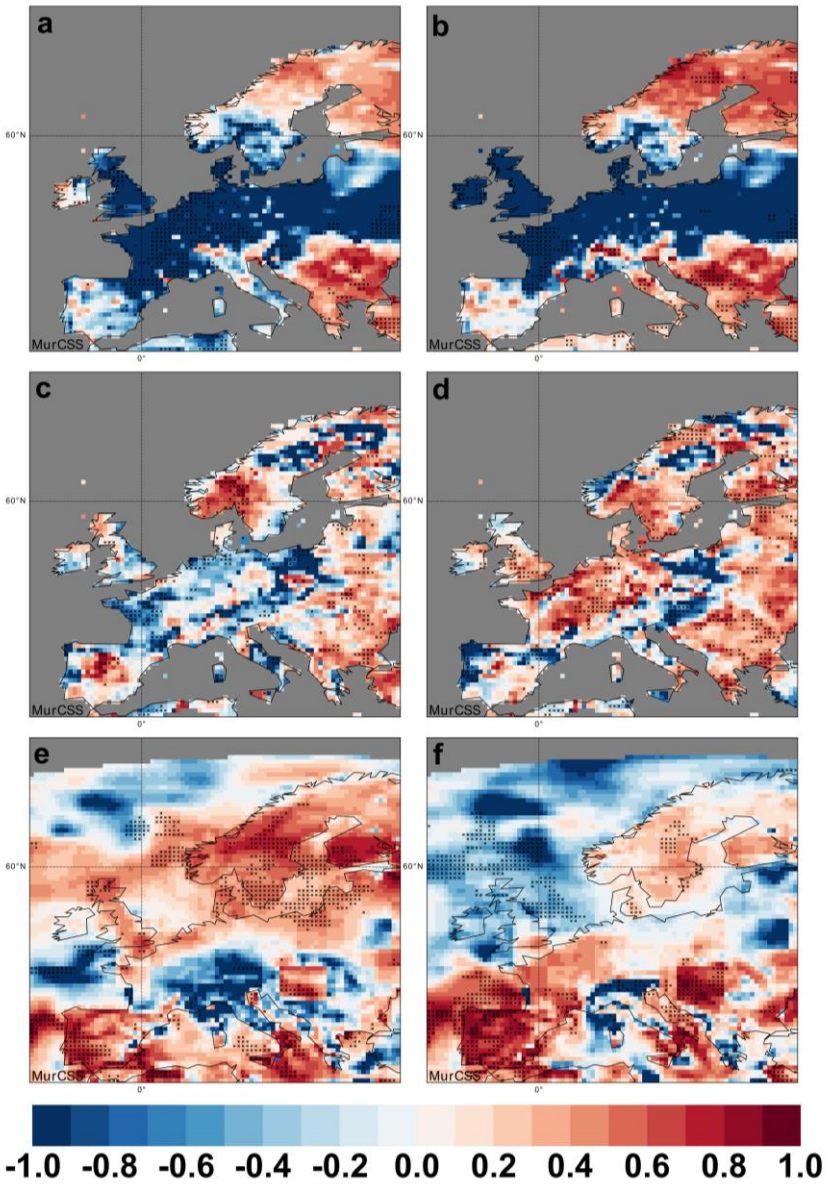

Figure 2: Spatial distribution of the MSESS for the multi-annual mean anomalies of lead years 1-5 for (a) temperature in CCLM_b0, (b) temperature in CCLM_b1, (c) precipitation in CCLM_b0, (d) precipitation in CCLM_b1, (e) wind speed in CCLM_b0, and (f) wind speed in CCLM_b1. As reference dataset we have used the uninitialized historical ensemble. The black dots indicate significant skill at the 95% level (bootstrapping).

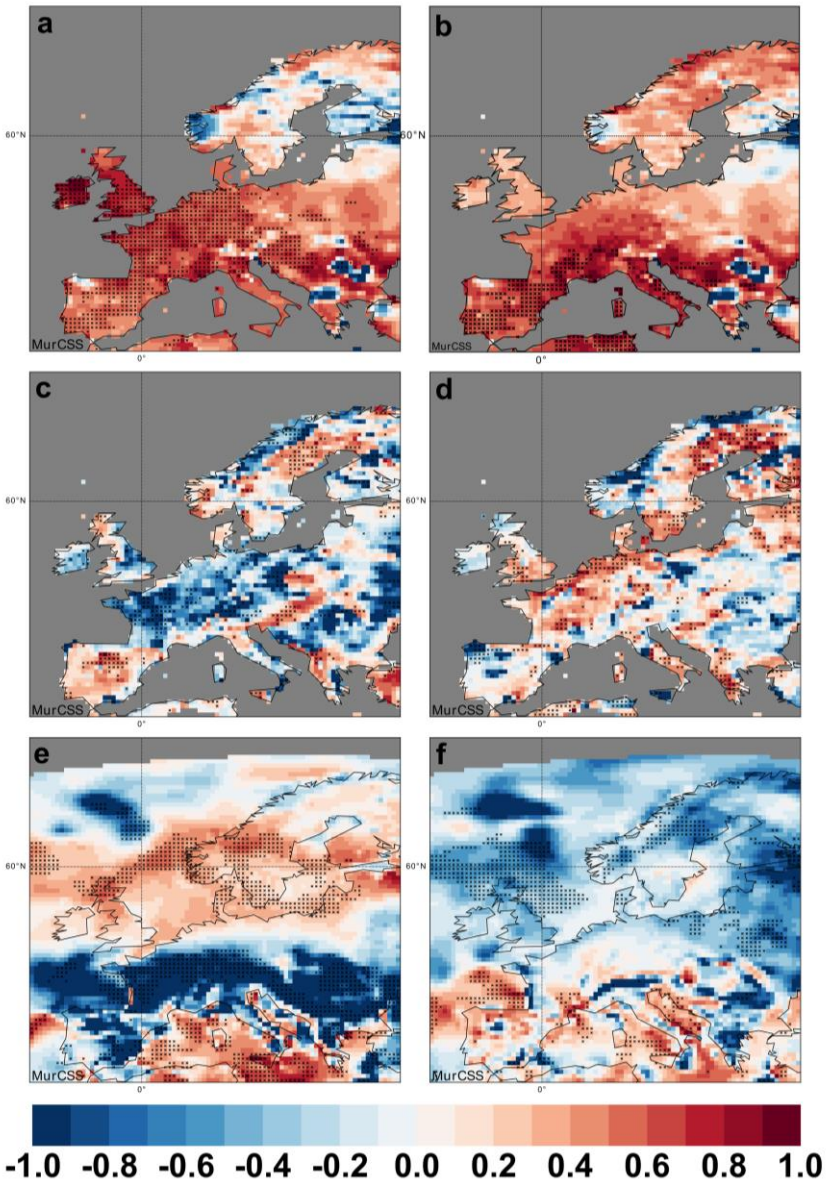

Figure 3: Spatial distribution of the MSESS for the multi-annual mean anomalies of lead years 1-5 for (a) temperature in CCLM_b0, (b) temperature in CCLM_b1, (c) precipitation in CCLM_b0, (d) precipitation in CCLM_b1, (e) wind speed in CCLM_b0, and (f) wind speed in CCLM_b1. As reference dataset we have used the climatology. The black dots indicate significant skill at the 95% level (bootstrapping).

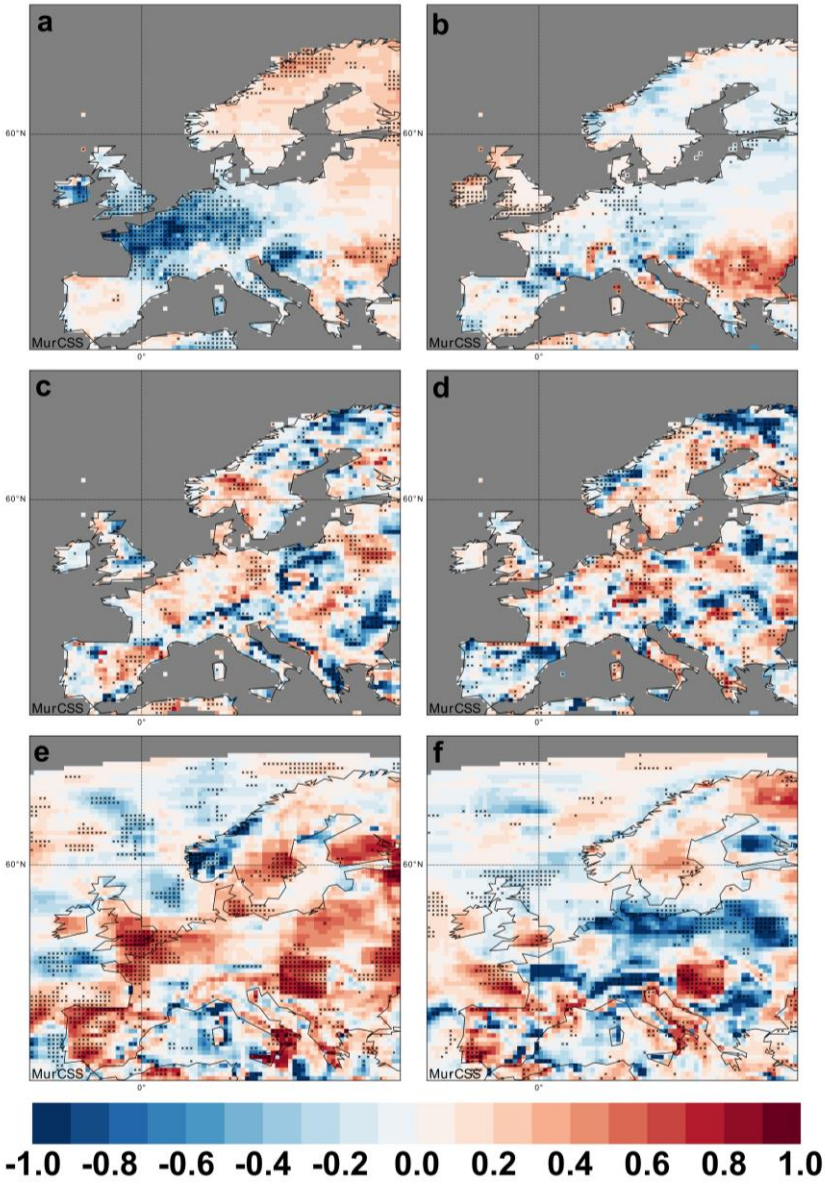

Figure 4: : Spatial distribution of the MSESS for the multi-annual mean anomalies of lead years 1-5 for (a) temperature in CCLM_b0, (b) temperature in CCLM_b1, (c) precipitation in CCLM_b0, (d) precipitation in CCLM_b1, (e) wind speed in CCLM_b0, and (f) wind speed in CCLM_b1. As reference dataset we have used the respective global MPI data. The black dots indicate significant skill at the 95% level (bootstrapping).

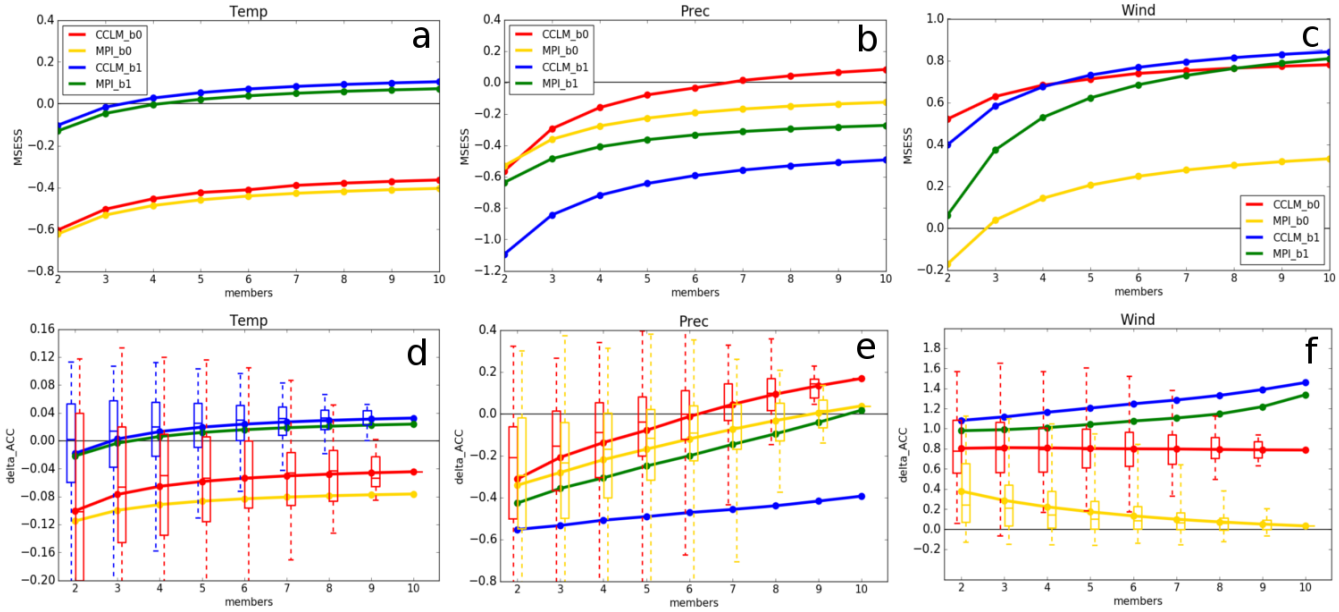

Figure 5: Skill scores for the multi-annual mean anomalies of lead years 1-5 of the CCLM_b0 (red), MPI_b0 (yellow), CCLM_b1 (blue), and MPI_b1 (green) ensembles depending on the ensemble size (x-axis, ranging from 2 to 10 members) over IP (cf. Fig. 1). MSESS for (a) temperature, (b) precipitation, and (c) wind speed; delta_ACC for (d) temperature, (e) precipitation, and (f) wind speed. In (d)-(f) box-whisker plots for the skill scores of all n-member combinations are shown. For MSESS and delta_ACC 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.

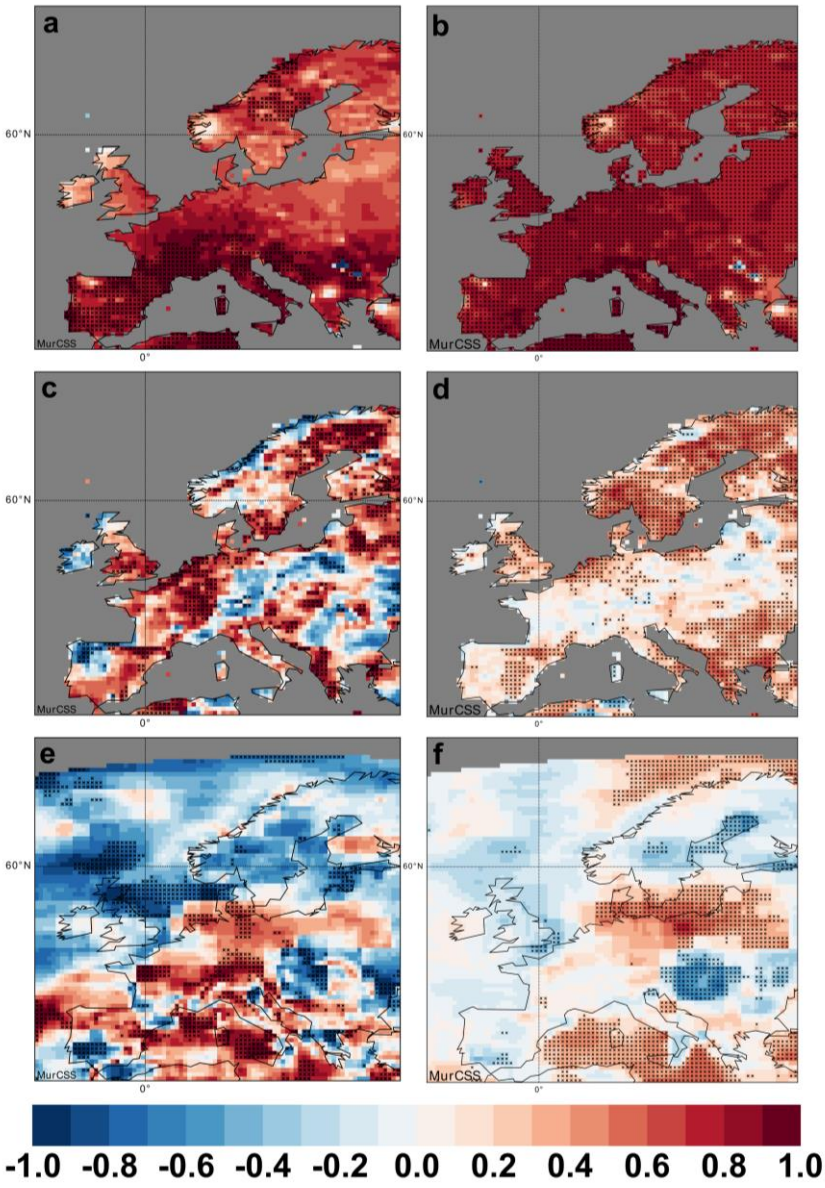

Figure 6: Spatial distribution of the ACC for the multi-annual mean anomalies of lead years 1-5 in MPI_b1 for (a,b) temperature, (c,d) precipitation, and (e,f) wind speed. For the left panels five start years (dec1960, dec1970, dec1980, dec1990, dec2000) have been used, while for the right panels all start years from dec1960 to dec2000 are taken into account. The black dots indicate significant skill at the 95% level (bootstrapping). For more details see main text.