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

Reviewer 2

Comment of the authors:

We thank the Reviewer for his/her thoughtful review and the specific suggestions. The Reviewer queried two major points: (i) the low number of starting dates used in our study (which has been discussed in detail in the original manuscript) and (ii) the detrending of the time series. As we generally agree with the Reviewer in both points and take the concerns seriously, we decided to include new Figures in our responses with respect to point (i) (see also answer to main concerns below), and to redo most of the calculations without detrending the time series as suggested by the Reviewer and replace the respective Figures in the revised manuscript. As the new calculations will impact the majority of our results quantitatively, a reformulation of most parts of the manuscript will be necessary. Therefore, in most of our responses to the specific comments (see below) we only stated which changes will be made, without giving the manuscript changes explicitly at this point in time. The point-to-point responses to all major and specific comments are marked in red.

Anonymous Referee #2
Received and published: 19 December 2017
The paper gives a preliminary assessment of regional decadal prediction skill over Europe based on a high-resolution regional model forced with boundary conditions obtained from the low-resolution, global MiKlip prediction system. I deem the analysis preliminary because the “development and prospects” of the downscaling system are being assessed at a rather early stage when only 5 hindcast start dates have been completed using the regional model. This is a serious shortcoming that calls into question the reliability of skill scores (computed from 5 data pairs) that are used throughout to make statements about the benefits of downscaling for various fields in various European regions.

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

Answer:
We agree that five starting years lead to “noisy” results. But we argue, that this noisiness affects chapter 4.1 the most, dealing with the skill distribution over Europe. The chapters 4.2 regarding the added value and chapter 4.3 regarding the ensemble size provide robust results even when using two times 50 simulations. Therefore, the major parts of the results are not affected strongly by the sample size issue.
To remedy the sample size and starting year issue we performed, as the reviewer suggested a comparative analysis of the skill estimates for the three variables addressed in the paper derived from a) starting years every 10 years (1960, 1970,..,2000) as in the original

manuscript and b) annual starting dates (1960-2005) for the global 10 member ensemble with MPI-ESM-LR baseline1. Baseline1is the only ensemble used in the paper which provides 10 members throughout the whole hindcast period. As reference we applied 10 member of the un-initialized "historical" ensemble with MPI-ESM-LR. The results (see figures for the correlation at the end of the document) show a general qualitative agreement, though of course not a quantitative one. As expected, larger sample size provides smoother skill estimates, less noisy than with the smaller sample size. But in general the findings regarding the regions with hindcast skill mentioned in the original manuscript are still correct for the extended ensemble. We offer to include this additional analysis in the paper to point out, how and where the smaller sample size affects the findings and that way putting them into perspective.

Another main concern is the use of detrending, which probably exacerbates the sampling issues (how well-defined is a trend computed from 5 data points?). There is no real need to detrend since you have an uninitialized ensemble that allows you to determine the skill improvement relative to the externally-forced signal (yes, pure ACC will be higher, but you can show delta(ACC), i.e. the change in ACC relative to the uninitialized ensemble).
The quality of the writing is decent, but not high, and there are numerous instances of poor English construction (some noted below). A thorough proofreading is in order if this is to be resubmitted.

A: We thank the Reviewer for this helpful comment. We agree that detrending is not necessary when using an uninitialized ensemble as reference. We therefore decided to redo most of the analysis of our study without detrending of the time series and include the new results in the revised manuscript. The figures attached at the end of this document (see also the response to the first main concern) show the ACC for the baseline1 generation and additionally the difference to the ACC of the uninitialized historicals as suggested by the Reviewer. The respective Figures without detrendeing for the other skill metrics will be included in the revised manuscript. Further, we will proofread the revised version of the manuscript as suggested by the Reviewer.

Specific Comments and Questions:
P2,L8: Here and throughout: "Yaeger" should be "Yeager".
A: Citation will be changed throughout in the revised version.

P2,L11: It's not clear what the point is of the "while few" construction. Are you contrasting the large number of studies focusing on global metrics with the relatively few studies focusing on storm tracks, etc? Please rewrite.
A: We agree with the Reviewer that this sentence is misleading. We will rephrase it in the revised version.

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

A: This sentence will be removed in the revised manuscript.

P3,L7-18: The motivation for the present work needs to be clarified, particularly since it is not at all clear how the present study differs from the closely related recent MiKlip studies that have just been cited (Kadow et al. 2016; Mieruch et al. 2014; Haas et al. 2016; Moemken et al. 2016).
A: As stated in line 4-6 on page 3, the closely related MiKlip studies are difficult to compare, as they use different skill metrics, pre-processing methods and downscaling approaches. The unique feature of our study is that we use the same methods/metrics not only for the

regional but also for the global prediction system, which enables us to give a more general assessment of the prospects of the MiKlip system with respect to near surface variables which affect human life most (temperature, precipitation, and wind speed). However, we agree that this sentence is not sufficient to emphasize our motivation. We will rephrase it in the revised version.

P3,L11: This question is poorly phrased. Do you mean “depend on” the trend or “derive from” the trend?
A: We agree that this question is misleading. We will remove the second part of the question.

P3,L15: This is a repetitive rephrasing of the questions just covered.
A: Paragraph in line 15-18 will be removed in the revised manuscript.

P3,L29: I don’t understand how ocean temperature and salinity can be nudged towards NCEP/NOAA reanalysis, since the latter is an atmospheric reanalysis.
A: The reviewer is correct. In baseline0 the ocean salinity and temperature anomalies were derived from a simulation with the ocean model MPI-OM forced with the NCEP re-analysis. We will change the sentence accordingly.

P4,L11: Not clear what is meant by “Analog to the global data”?
A: To avoid misinterpretation, we will change the text:
“The experiment includes downscaled hindcasts for dec1969, dec19790, dec1980, dec1990, and dec2000, with ten members per decade (hereafter CCLM_b0 and CCLM_b1). The regional ensembles therefore consist of the same time series like the global ensembles MPI_b0 and MPI_b1.”

P4,L14: Replace “are” with “is”.
A: Will be replaced in the revised manuscript.

P4,L16-19: You already introduced the ERA-driven CCLM simulation in the first line of this paragraph, so consolidate your sentences into one brief description.
A: We will change the first sentence of this paragraph:
“To evaluate the performance of both the global MPI-ESM and the regional CCLM hindcasts, reanalysis and observational datasets are used for verification.”

P4,L21: I don’t understand what you mean by “uninitialized model simulations started from historical CMIP5 runs”. Do you mean downscaled simulations that can be considered “uninitialized” counterparts to CCLM_b0 and CCLM_b1? Do you mean “preindustrial CMIP5 runs”?
A: We agree that this sentence is misleading. We will change it in the revised version:
“To address this issue, uninitialised historical CMIP5 runs are usually considered …”.

P4,L32: Replace “the natural variability” with “natural variability”. Why use linear detrending to isolate natural variability when you have just introduced an uninitialised ensemble that can be used to quantify the skill associated with external forcing?
A: The forecast skill of a decadal prediction system may origin from two different “processes”: a realistic prediction of the long-term trend and a suitable forecast of peaks on inter-annual timescales due to natural variability. To isolate the forecast skill for anomalies on inter-annual time scales, we originally detrended **all** datasets used in this study (as stated in the first paragraph of section 3.1), i.e. not only the hindcasts and the observations, but also the uninitialized historical runs. However, we agree with the Reviewer that detrending is not necessary when we use an uninitialized ensemble that allows us to determine the skill improvement relative to the externally-forced signal (see Reviewer’s major comments). Hence, we decided to redo most of the analysis without detrending in the revised version (see also our answer to main comments). We will further replace “the natural variability” with “natural variability”.

P5,L14: I think you mean “post-processed time series”.
A: No, we mean pre-processed here, as they are processed before they are analysed by using different skill-metrics.

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

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

P5,L25-: What are F(y) and Fo(y)? Please explain the CRPS equation. What exactly is CDF and how is it computed?
A: The CRPS is defined as the quadratic measure of the discrepancy between the forecast cumulative density function (*F)* and the observed cumulative density function (*Fo)* of a variable *y*. The cumulative density function (CDF) of a real-valued variable *y* is defined as: CDF(y) = P(y ≤ t),
where *P* is the probability that the variable *y* has a value of less than or equal to *t*.
We will add this information to the revised manuscript.

P6,L23-P7,L2: This is repetitive.
A: This paragraph will be removed in the revised manuscript.

Fig 2: Suggest using a nonlinear scale for MSESS, such as -9 to 0.9 as in Shaffrey et al. (2016, doi:10.1007/s00382-016-3075-x), because this metric is not symmetric about 0 in terms of relative improvements in MSE. Please clarify that these are for annual mean (ie, not seasonal mean) anomalies.
A: Following the Reviewers suggestion, we will use a nonlinear scale for Fig. 2 in the revised version and clarify that skillscores are for annual mean anomalies.

P7,L7: I presume the detrending has been performed similarly for observations and for the uninitialized historical runs? This isn’t explicitly mentioned.
A: In the first paragraph of section 3.1 we stated that all datasets “are pre-processed in an analogous manner”. However, to avoid misunderstanding we will add this information here again.

Table 1: What is the meaning of “The uninitialized historical ensemble has been used as reference dataset”, given that this is a table of ACC scores? Am I correct that this table displays correlations computed from 5 data points (corresponding to the 5 start years)? Clearly the externally-forced trend is important and so this table should include ACC scores for the uninitialized historical runs for comparison.
A: We thank the Reviewer for this note. Up to now, no information of the uninitialized historical ensemble is included in Table 1. In the revised version we will include ACC scores for the uninitialized runs following the Reviewer’s suggestion. As we use multi-annual means, correlations are actually computed from only 5 data points, which may be problematic as already criticized by the Reviewer. However, as shown in the responses to the main comments and in the attached figures, using 5 starting dates instead of yearly initialized hindcasts has mainly quantitative effects.

P7,L11: What is the meaning of “increases” ˇTrelative to uninitialized or relative to detrended?
A: The MSESS for the datasets with trend increases relative to the detrended time series. We will clarify this in the revised manuscript.

P7,L34: I would say Figure 2 shows more than a “slight shift”.
A: This is indeed a too strong generalization of the results. Discrepancies between the two hindcast generations are rather small for temperature, but can be quite large for precipitation and wind speed, depending on the region. We will clarify this in the revised version.

P8,L2: Here and elsewhere delete “exemplary” as it is not being used properly.
A: Following the Reviewer’s suggestion we will delete exemplary.

P8,L4: It’s curious that Fig 2e agrees so well with Fig 3a, but Fig 2f is so different from Fig 3b. Can you offer any explanation? In my mind, it calls into question the significance of skill scores computed from 5 data points.
A: It is difficult to find an explanation for this issue. Aside from statistical reasons this might also be related to the detrending of the data. However, as we intend to replace this Figure by new ones obtained from calculations without detrending (see also response to main concerns), there is no point to speculate at this point in time. We will carefully check our new results for such discrepancies.

P8,L25-29: This discussion begs the question of why you are doing any detrending at all (see comment above)? The purpose of the uninitialized ensemble is precisely to allow you to discriminate between greenhouse-gas induced variability (including trends) and natural variability (including AMO-related trends). Detrending is confusing matter sâˇAˇ Tjust compared initialized to uninitialized skill.
A: Again, we agree with the Reviewer in this point and will redo most of the calculations without detrending.

P9,L8: Change “whereas” to “and”.
A: Will be changed.

P9,L9-11: This incomprehensible sentence needs a rewrite.
A: We will rephrase this sentence in the revised manuscript as it is indeed incomprehensible.

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

P11,L24-28: This is because you are doing bootstrapping without replacement; if you allow replacement, then the spread does not necessarily diminish with ensemble size.
A: As stated in line 2 of page 11 of our manuscript our aim was not to do a bootstrapping, but to do permutations over all useful ensemble combinations. In our opinion the individual n-member ensembles should contain each ensemble member only once. Otherwise, the 10 member ensemble may in an extreme case consist of 10 times the same member, which in our opinion makes no sense. However, it is correct that the permutation without replacement results in a decline of the spread with increasing number of members. We will add this information to the paragraph in the revised manuscript.

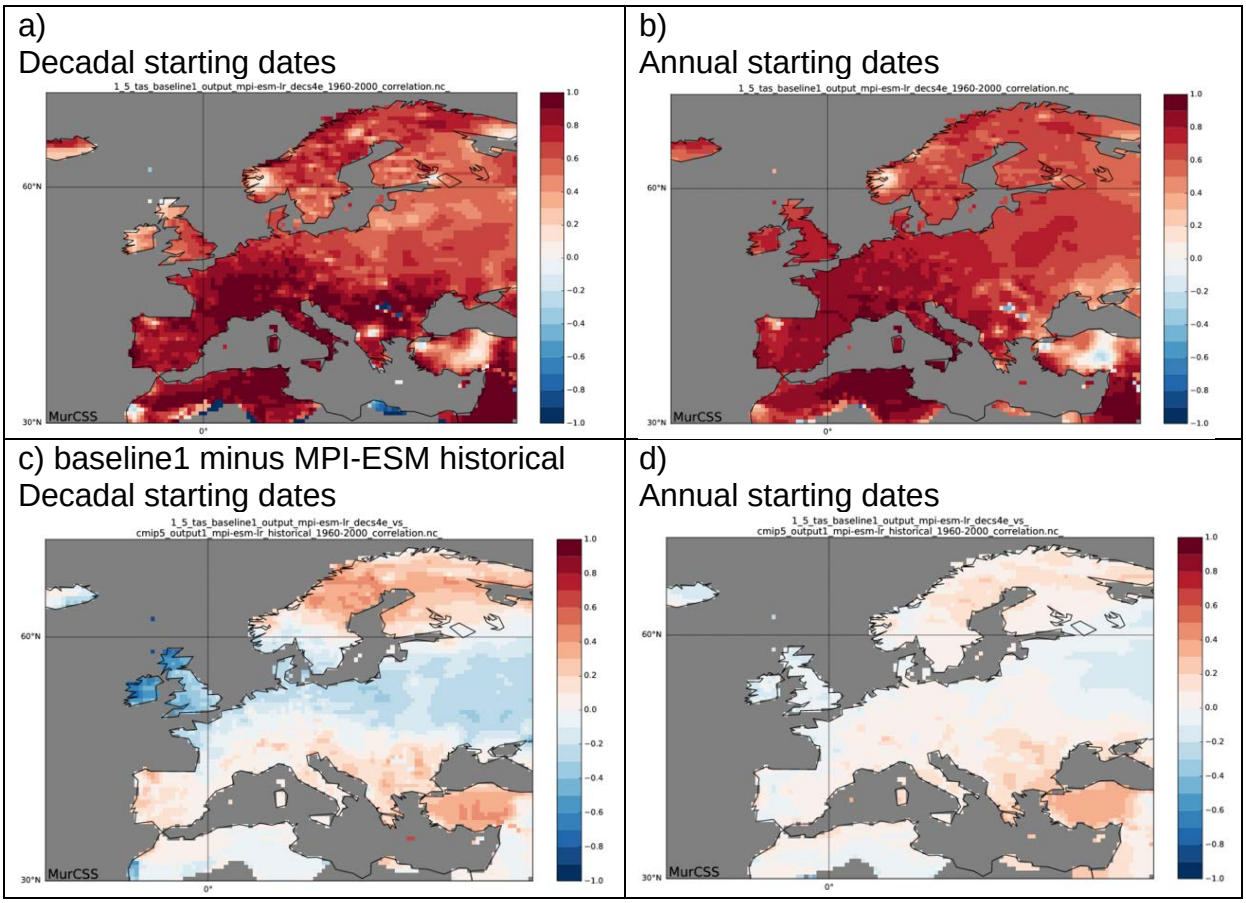

Fig A1: (a,b) **Temperature correlation MPI-ESM-LR baseline1,** 10 members, lead-times year 1-5, observation: E-OBS. (c,d) Correlation baseline1 minus MPI-ESM-LR historical.

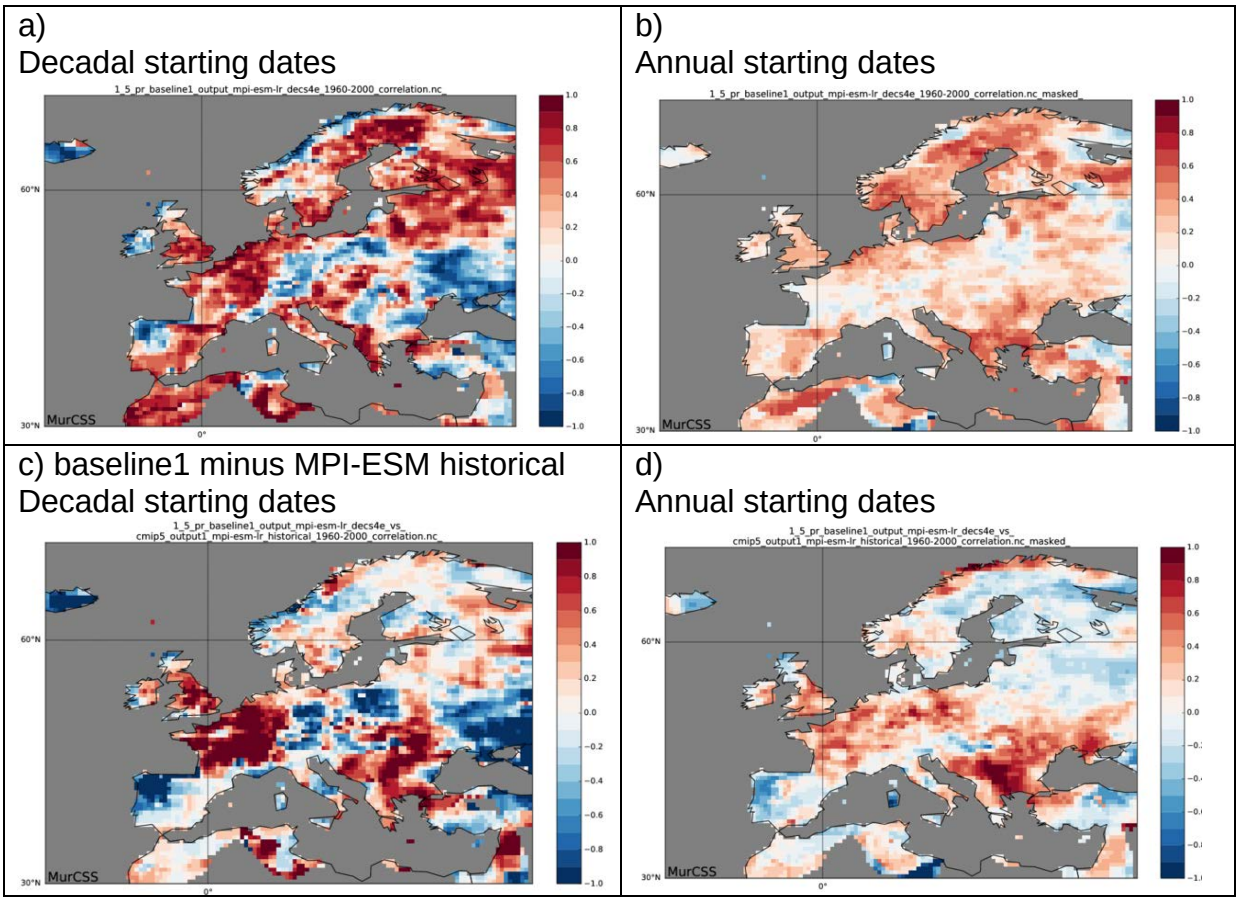

Fig A2: (a,b) **Precipitation correlation MPI-ESM-LR baseline1,** 10 members lead-times year 1-5, observation: E-OBS. (c,d) Correlation baseline1 minus MPI-ESM-LR historical.

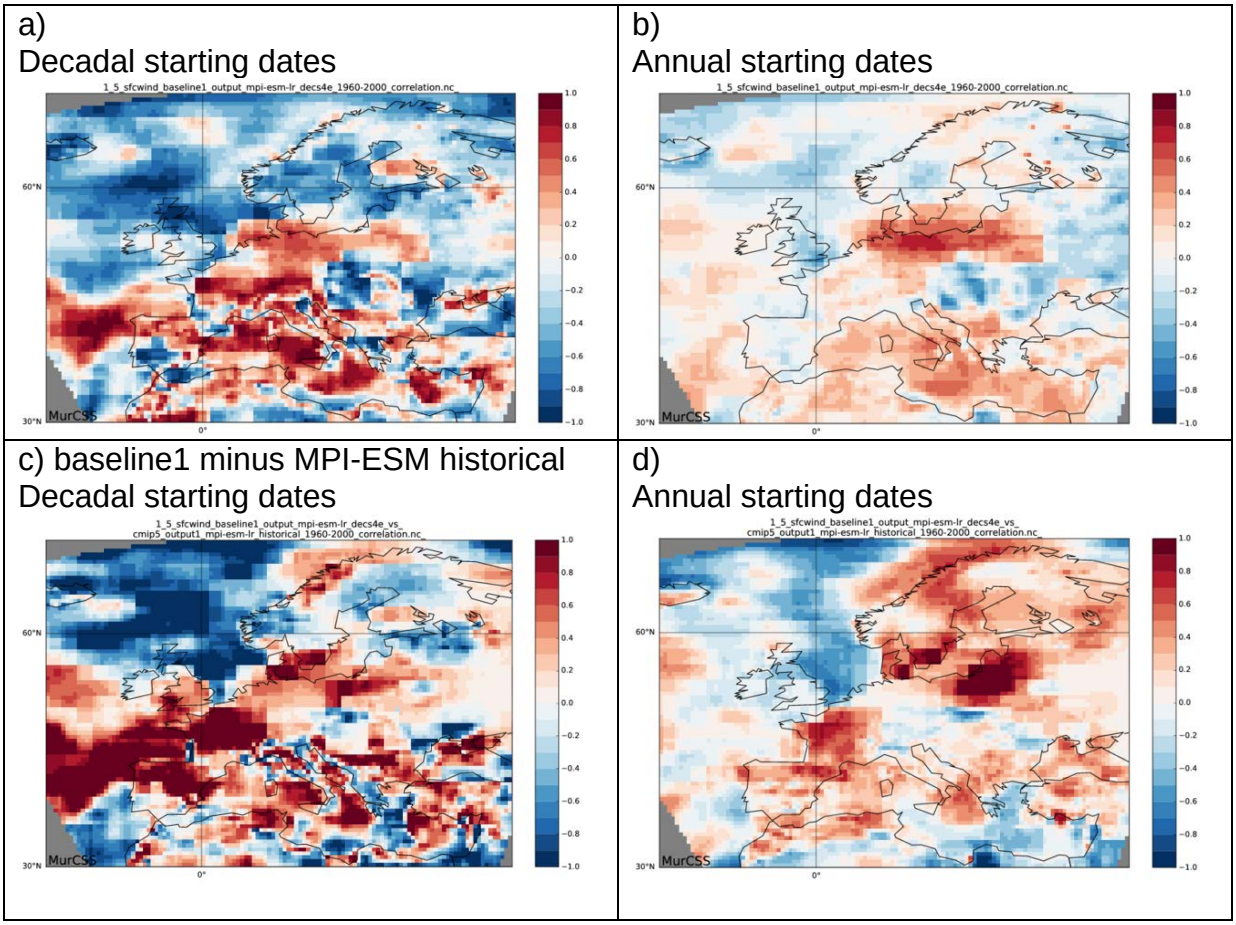

Fig A3: (a,b) **Surface wind correlation MPI-ESM-LR baseline1,** 10 members lead-times year 1-5, observation: CCLM ERA 0.22°. (c,d) Correlation baseline1 minus MPI-ESM-LR historical.

Interactive comment on Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2017-70, 2017.