NUKLEUS – A first kilometer-scale convection-permitting multi-model climate ensemble for Germany: Characteristics of the historical simulations 1961–1990

Final Author Response

Response to Reviewer Comment (RC) 1: Sven Kotlarski

Braun et al. present a thorough evaluation of the GCM-driven NUKLEUS ensemble over Germany. The ensemble consists of nine simulations, carried out by three different regional climate models run at convection-permitting resolution of 3 km and driven by three different global climate models in a double nesting approach. The evaluation targets the historical reference period 1961-1990 and focuses on 2m air temperature, precipitation and 10m wind speed at annual, seasonal, daily and sub-daily scale. The paper is accompanied by 2-3 further submissions that analyse further aspects of the same ensemble or that evaluate the reanalysis-driven counterpart of the present GCM-driven nine-member ensemble. A range of evaluation metrics that are applied over six sub-domains are considered. The gridded German HYRAS dataset as well as ERA5-Land serve as evaluation reference datasets, hence including some notion of reference data uncertainty. The evaluation yields, in general, satisfactory results that are located in the range of previous RCM evaluations or partly even improve upon them. A persistent warm bias at least over the southern part of the analysis domain (Southern Germany, pre-Alps) and a general wet bias are identified. These mean biases are successfully (and by design) reduced by the application of a quantile-mapping based bias adjustment. It is also shown that, despite the successful mitigation of mean and distributional biases, event-based spatial fields might remain spurious after bias adjustment.

The manuscript is, in general, well structured and well illustrated. The analyses are carefully designed and explained for most parts. The paper is of a rather descriptive nature, but still tries to include explanations and interpretations where possible. The number of figures is large, but still in an acceptable range in my opinion. Side analyses are outsourced to a supplementary document. The results obtained are carefully and successfully placed into the context of previous works. Overall, I consider the scientific value of the study as extremely high. The work could almost serve as a role model for a careful model evaluation that should always precede the actual use of climate model ensembles. Still, the manuscript might profit from a number of additions and/or modifications, some of them are listed below. Including them is, however, not a must in my opinion. The manuscript is basically valuable as is. A couple of further minor issues are also listed below.

My suggestion is to consider a publication of the manuscript after minor revisions.

With kind regards.

Sven Kotlarski

Answer: Thank you for the very constructive feedback on our manuscript! Please find our answers below. Based on your and the second reviewer's suggestions we decided to implement the following

changes to the manuscript:

- Change main references for all hourly data and for daily near surface wind from ERA5-Land to TRY and RADKLIM
- Sharpen comparison with NUKLEUS ERA5-driven evaluation simulations (Sieck et al., 2026) with a focus on biases originating from RCMs and the forcing data
- Separate analyses by seasons, where appropriate, also for a better comparability with NUKLEUS ERA5-driven evaluation simulations by Sieck et al., 2026
- Strongly shorten section 5 (comparison of the NUKLEUS ensemble with other ensembles) and merge into discussion
- Strongly shorten section 6.2 (event-based analysis of bias correction) and merge into 6(.1)

IDEAS FOR IMPROVEMENTS/EXTENSIONS

Reference period 1961-1990 and trends since then: The authors correctly motivate their choice of reference period by WMO norms for monitoring long term climate change. However, the considered period dates back already more than 30 years, and most accompanying works apply more recent reference periods (especially the current 1991-2020 period). One advantage is the typically better availability and quality of reference datasets in later periods. But also, future climate change is often referenced to some "present-day" climate, particularly in climate impact assessments. Also the entire chapter 5 seems to have some inconsistency here (1961-1990 results from the present work are compared to 1981-2010 results for EURO-CORDEX or 1991-2020 results for KIT-KLIWA). Hence my question, why the authors did not choose a more recent reference period. A fair representation of the 1961-1990 climate does not necessarily imply a fair representation of the more recent climate, in particular since several recent works have documented problems of at least the CMIP5-driven EURO-CORDEX ensemble to represent historical temperature trends. Connected to this, the representation of actual historical temperature TRENDS is not covered by the analysis presented, but could add considerable value.

Answer: We agree that there are good reasons to consider a later reference period. However, due to computational constraints given the relatively large size of the ensemble, the 3km simulations were performed for specific time slices of 30 years as mentioned in section 2.1. Since one of the main goals of the NUKLEUS project is to transfer knowledge from science to practice in terms of climate adaptation, these time slices were selected in a co-design process with stakeholders and users of downstream applications within the pilot regions of the RegIKlim funding measure in the design phase of the NUKLEUS ensemble. Based on that exchange, the reference period 1961-1990 was chosen. Consequently, choosing a different period is unfortunately not possible. We will include a statement in the revised version that describes this better.

We also agree that analyzing the simulated trends (esp. of the 12 km simulations due to their transient nature) is of interest and helpful to put the 3 km simulations into context. Yet, the focus is here on the 3 km simulations as they are in the main focus of the RegIKlim stakeholders. As a compromise, we will include time series in the supplement comparing the NUKLEUS 3km and 12km, the forcing

GCMs and HYRAS for the period covering 1950 to 2020 (see Fig. 1 below) and add a short paragraph on that in the main text.

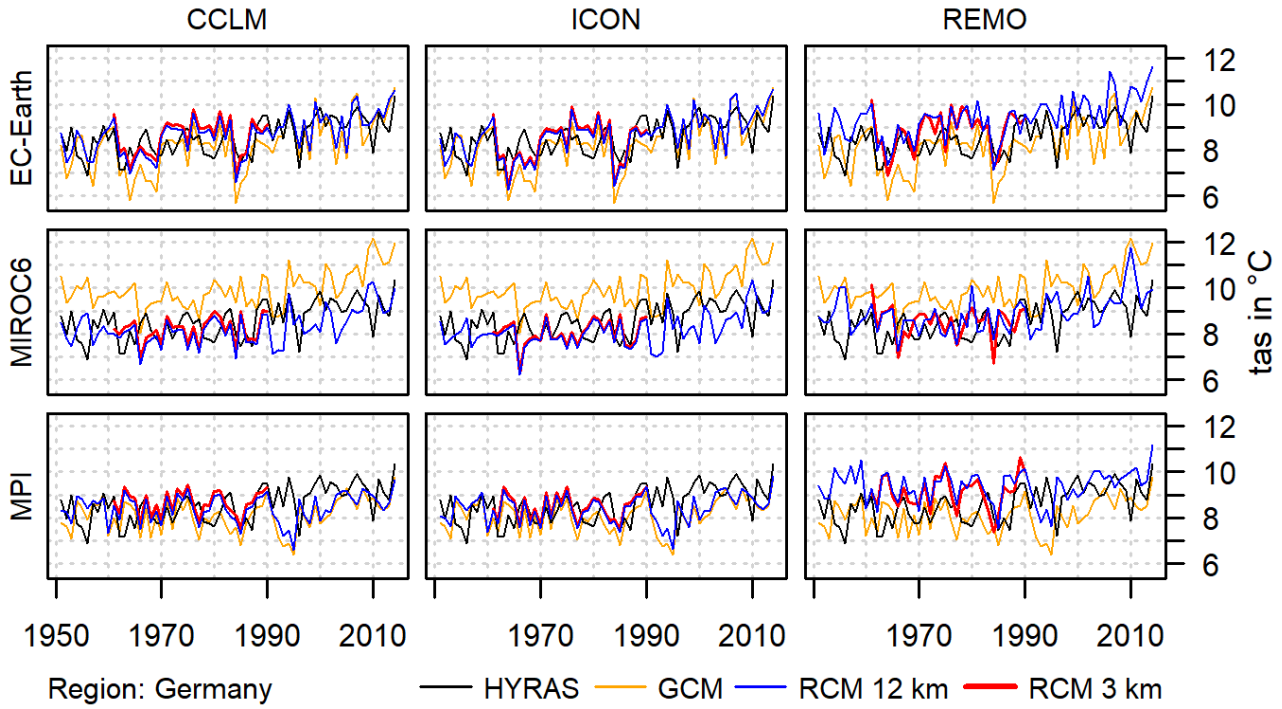


Figure 1: Time series of annual mean 2m temperature averaged over Germany for the three RCMs (ICON, CCLM, REMO) driven by the three GCMs (EC-Earth, MIROC6, MPI) at both resolutions (blue: 12km; red: 3km), the original GCM data (orange), and HYRAS observations (black) for the historical period.

Focus on aspects of added value of the convection-permitting scale and more direct comparison to the intermediate EUR-12 ensemble: The presented analysis is "standard" in many aspects and does only partly target aspects where the added value of convection-permitting modeling can be assumed (extremes in strong topography, spatial variability in strong topography, representation of high elevations, summer precipitation extremes and diurnal cycle, etc.). One could shift the analysis much more into this direction, for instance by assessing the representation of the precipitation diurnal cycle separately for each season (instead of the annual man diurnal cycle). Also, the comparison to the performance of the intermediate EUR-12 ensemble (first nesting step) is obviously covered by an accompanying paper. It would be highly interesting to include it more explicitly here in order to identify the added value of the (still costly) km-scale.

Answer: We agree that the analyses are largely standard, reflecting our aim to provide a comprehensive assessment of the climate represented by the NUKLEUS 3 km ensemble in its current form. We consider the benefits of convection permitting simulations compared to coarser ones already well documented in the literature and will refer to those more extensively in the revised manuscript. At the same time, we fully acknowledge that the NUKLEUS ensemble offers great potential to further investigate the added value of convection permitting climate simulations. Yet, this is currently ongoing work. Still, we decided to cluster some of the analyses by season, in order to provide deeper insight into the quality of the data (see as example Fig. 2).

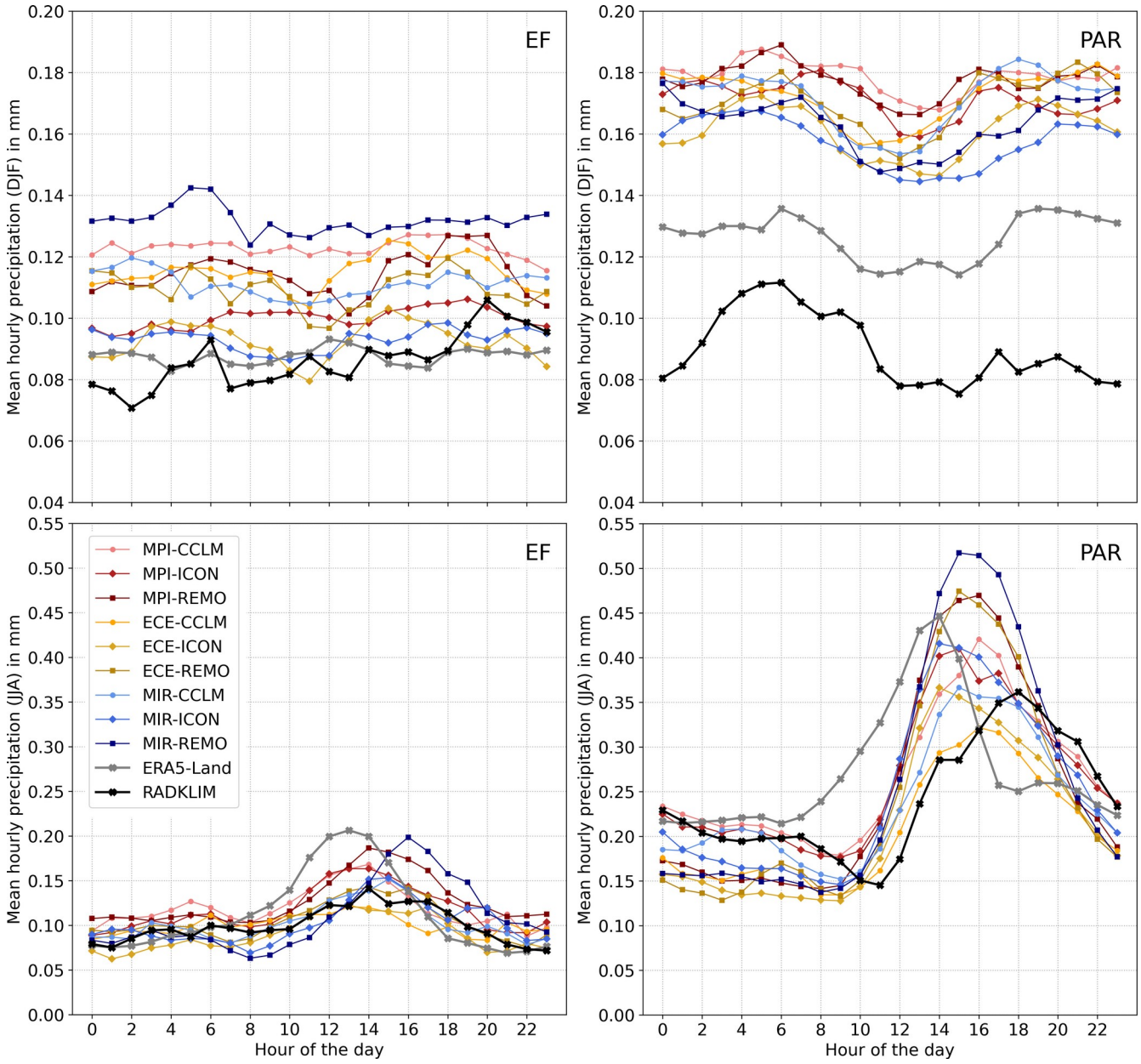


Figure 2: Diurnal cycles of the pilot regions EF (left) and PAR (right) for winter (DJF, top) and summer (JJA, bottom) for all NUKLEUS 3km simulations and ERA5-Land in the historical period 1961-1990, and RADKLIM in the evaluation period 2005-2014.

MINOR ISSUES

Line 2: "for Germany" -> the model domain also covers adjacent areas that might be worth to be analyzed. I suggest to rephrase to "for Germany and adjacent areas".

Answer: We agree. The text will be adjusted.

Lines 2-4: This objective is a little too strong in my opinion. At least, it cannot be answered by this study alone, as the answer to the question also depends on the nature of the future scenario simulations (sampling of regional response space, possibility to properly quantify uncertainties, etc.).

Answer: We agree that the scope is too wide for this particular manuscript and will adjust accordingly to: "The main goal is to examine to what extent these high-resolution simulations reproduce the observed climatic conditions over Germany, which is a basic requirement to derive actionable climate information."

Line 6: "spatio-temporal and statistical representation" -> also spatio-temporal statistics are a statistical measure. What you probably mean is "spatio-temporal and distributional representation".

Answer: Thank you for pointing out the duplication. We will rephrase this to: "The evaluation focuses on spatio-temporal patterns and statistical distributions of basic meteorological variables…"

Line 24: "programme" instead of "program".

Answer: True. Will be adjusted.

Line 45: "and its Flagship Pilot Studies" -> The FPSs are on overall CORDEX initiative and not specific to EURO-CORDEX (as implied by this sentence).

Answer: Valid point, we will adjust the text accordingly.

Fig. 1: This figure shows the full analysis domain, but not the full regional model domain. For the full modelling domain the reader is referred to Sieck et al. , which is a little unfortunate. Consider to show the full model domain also here (it seems to be only slightly larger), similarly to Sieck et al.

Answer: We assume the reviewer is referring to Fig. 2 in the manuscript. We will replace this figure with one showing the entire model domain.

Line 108: "simulations of THE historical reference period".

Answer: Will be adjusted.

Section 2.1.2: It would be interesting to know whether or not a retuning of the RCMs when switching from 12 km (intermediate nest) to 3 km has been carried out or not. Is it only the deep convection that has been switched off or have further changes been applied?

Answer: For each of the RCMs, different scale dependent parameter setups from earlier experiences and projects were used on 12 km and 3 km, respectively. So yes, further changes have been applied, except switching off the parameterization for deep convection. We provide the most relevant details below. In the revised version of the manuscript we will line out the differences to the downscaling steps described in the companion NUKLEUS RCM evaluation manuscript by Sieck et al. (2026), in order

to avoid unnecessary duplication of information.

- For REMO, this implies the change from REMO to REMO-NH and the usage of a set of parameters as obtained from the EUCP project.
- For ICON, in the convection permitting simulations at 3 km resolution, deep and mid-level convection parameterizations are switched off, so that only the shallow convection part is remaining. Furthermore, the scale dependent ICON setup (as recommended by DWD) comprises the usage of an extended microphysics scheme on 3 km, more atmospheric levels, i.e. 65 (3 km) instead of 60 (12 km), different top-height, i.e. 22 km (3 km) instead of 23.5 km (12 km), and non-orographic gravity wave drag (inwp_gwd) switched off in 3 km. With these parameter sets, tests were executed, but no retuning was performed.
- For CCLM, numerous scale-dependent parameter adjustments are applied, most notably in the cloud parameterization, e.g., extending the microphysics scheme by an additional cloud-ice class for graupel, and in the turbulence parameterization.

Lines 168-170: The SST and lake treatment and the soil initialisation information is only provided for REMO, but not for CCLM and ICON. I'd suggest to either remove it from here or to also provide this information for the two other models.

Answer: As the main description of the RCM setup is given in detail in Sieck et al. (2026), to which we refer to, we agree on removing this detail here.

Lines 185-186: This statement is somehow critical as it implies that the step from ERA5 to ERA5-Land is the main reason for differences between the simulations and ERA5-Land. Many more reasons could actually be responsible. The discussion of differences between ERA5-Land and HYRAS would be more relevant in this context.

Answer: We agree that a comparison of the different reference datasets is more appropriate here. Since we plan to rely our model assessment for near surface wind on all time scales and the other variables on hourly basis stronger on TRY and RADKLIM, a comparison of ERA5-Land to those datasets instead of HYRAS appears more appropriate to us. We will adjust this in the revised version of the manuscript.

Lines 208-209: Very unclear.

Answer: We assume this refers to the sentence "Additionally to the pdfs, the bias relative to the reference data is examined". We agree that it is not clear what was meant. We will change it to "Additionally to the pdfs, **the spatial distribution of** the bias relative to the reference data is examined."

Lines 295-300: Wouldn't this bias pattern also imply an underestimation of the diurnal temperature range/DTR?

Answer: Yes. We can confirm that DTR is underestimated in simulations compared to HYRAS (not shown in the manuscript). We will include this conclusion in the revised version.

Lines 301-303: Some analysis or at least a reference should be provided for this statement.

Answer: The information that "this bias is strongly influenced by the land surface scheme" is not correct and will be removed. The distinct local warm and cold biases coincide with urban areas and lakes and originate from the corresponding surface properties and parameterizations. We will adjust this statement in the revised manuscript.

Figure 7 (and further): The yellowish (ECE) and reddish (MPI) lines can hardly be distinguished. Any chance to use color groups that are better separated from each other?

Answer: We agree that sometimes it is hard to distinguish between the lines. However, it is difficult to find three groups of colors (one for each GCM) with three variations (one for each RCM) that fits best. As a compromise, we will increase the marker size, which already is a second level of separation on the RCM level, and will introduce a third level using different line styles for the three RCMs. Hopefully, in most cases, this will improve the visibility.

Line 375: I think this is only valid for the monthly scale but, for instance, not for the annual scale (cf. Fig. 3). Here, the RCM seems to be at least as important. Better specify what you mean.

Answer: We agree that it is important here to point out the relevant time scale. We will adjust the statement accordingly.

Lines 391-396: This discussion somehow implies that the pronounced topography in PAR is responsible the pronounced annual precipitation cycle. Or, the other way round, that a flat topography leads to a flat annual cycle. This is surely not the full picture. I suggest to rephrase this part.

Answer: We agree that the statement is too simplistic and that there are definitely more drivers behind. We will rephrase this paragraph.

Lines 437-439: This section is a little too simplistic I'd say. To my knowledge, ERA5 does not assimilate precipitation station observations, but gridded precipitation products. This is then inherited by ERA5-Land.

Answer: You are totally right. For precipitation, ERA5 uses the short-range forecasts of the IFS cycles, ground-based observations are not directly assimilated, and other gridded products like satellite measurements are only available and used for the more recent years. However, the IFS system itself is using an assimilation system for observations (Lavers et al., 2022; https://doi.org/10.1002/qj.4351).

ERA5-Land is then a simple interpolation to a higher resolution as mentioned in the data section. Nevertheless, we agree that the given statement is too simplistic in that sense and that there are multiple reasons for the elaborated differences between ERA5-Land, HYRAS, and NUKLEUS. We will adjust and rephrase this paragraph accordingly in the revised version of the manuscript.

NUKLEUS – A first kilometer-scale convection-permitting multi-model climate ensemble for Germany: Characteristics of the historical simulations 1961–1990

Final Author Response

Response to Reviewer Comment (RC) 2: Dominic Matte

Review of "NUKLEUS - A first kilometer-scale convection-permitting multi-model climate ensemble for Germany: Characteristics of the historical simulations 1961–1990" by Christoph Braun et al. This manuscript is part of a three-paper series and focuses on the analysis of historical GCM-driven simulations produced using a double-nesting approach (GCM =>12-km => 3-km grid spacing). The simulations are performed over Germany, with the analysis conducted over six subregions. Model performance is evaluated against two reference datasets, HYRAS and ERA5(-Land), with a focus on seasonal and mean biases, variability, annual and diurnal cycles, and the overall precipitation distribution. The manuscript is generally well written and provides a detailed description of the main biases. However, the choice of reference datasets does not always help clarify the origin of these biases, particularly because of the relatively coarse resolution of ERA5(-Land). In my view, the manuscript could therefore be improved by sharpening the focus of the analysis and shortening some of the comparisons. Therefore I recommend accepting this paper for publication after what I would consider a minor revision.

Answer: Thank you for the critical and constructive feedback on our manuscript! Please find our detailed answers below. Based on your and the other reviewer's suggestions, we decided to implement the following noteworthy changes to the manuscript:

- Change main references for all hourly data and for daily near surface wind from ERA5-Land to TRY and RADKLIM
- Sharpen comparison with NUKLEUS ERA5-driven evaluation simulations with a focus on biases originating from RCMs and the forcing data
- Separate analyses by seasons, where appropriate, also for a better comparability with NUKLEUS ERA5-driven evaluation simulations
- Strongly shorten section 5 (comparison of the NUKLEUS ensemble with other ensembles) and merge into discussion
- Strongly shorten section 6.2 (event-based analysis of bias correction) and merge into 6(.1)

General comment:

Throughout the study, the authors repeatedly note that ERA5(-Land) is not well suited for comparison with CPM

simulations because of its relatively coarse resolution. This makes several of the comparisons difficult to interpret: when differences are subsequently attributed to the resolution of ERA5(-Land), it is unclear what is actually learned from the comparison.

I therefore wondered why the CPM simulations were not also evaluated against station observations, which would provide a more direct assessment and help separate model deficiencies from differences related to the reference dataset. Given the length of the paper, I would have appreciated seeing some results along these lines. However, I recognize that adding a station-based evaluation would represent a substantial amount of additional work.

Answer: We selected ERA5-Land as a reference, since it is the only gridded dataset known to us, that offers hourly data for all variables investigated here and data for near surface wind for the historical reference period. For the period 1961 to 1990 hourly station data for precipitation is not available at all (only starting from 1995). Furthermore, from our point of view within this period the availability of hourly station data for temperature and near-surface wind is too sparse to meaningful evaluate (the representation of variability within) the NUKLEUS data. Thus, we had made the decision for ERA5-Land and against the use of station data. However, we acknowledge that referring to the shortcomings of ERA5-Land is not satisfying.

Therefore, we decided to include the TRY wind data for the period 2005-2014 (evaluation period in Sieck et al.) into the wind-related plots (annual/diurnal cycles, pdfs) of the manuscript. Similarly we will use RADKLIM (2005-2014) as the primary reference for hourly precipitation and TRY (2005-2014) as the primary reference for hourly near surface air temperature (considering the offset due to warming between the historical and evaluation period). The bridge between the periods will be based on ERA5-Land and will be provided in the supplemental material.

Regarding near-surface wind we had already presented TRY in comparison to ERA5-Land in the supplemental material of the manuscript and found negligible differences in wind speed characteristics between the evaluation and historical period. From our point of view this qualifies the usage of TRY data from the evaluation period as the primary reference for the historical period. We will provide similar arguments for hourly near-surface air temperature and precipitation (see Figure 3 below).

As an additional benefit using the same reference on both manuscripts also facilitates the comparison of the historical and evaluation simulations and thus the attribution of origins of the biases to RCMs and the forcing data.

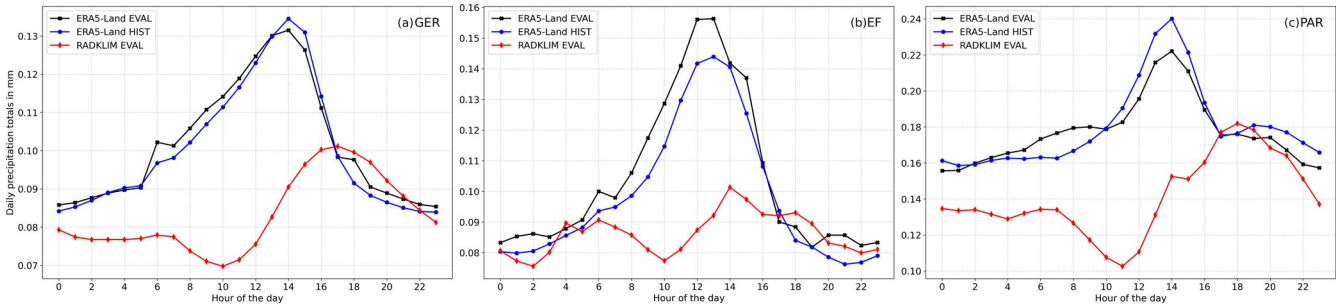


Figure 3: Comparison of the diurnal cycles of precipitation between ERA5-Land in the historical period 1961-1990 (blue), ERA5-Land in the evaluation period 2005-2014 used by Sieck et al. (black), and the RADKLIM data in the evaluation period (red).

An alternative would be to rely more strongly on the companion evaluation paper, assuming that it provides a sufficiently robust evaluation against observations and not only against ERA5(-Land). In that case, the present paper could briefly summarize the structural biases identified in the evaluation paper and focus primarily on biases introduced by the lateral boundary conditions (LBCs). This would avoid much of the discussion about scale mismatch, since the GCM-driven simulations could be compared directly with the corresponding evaluation runs.

I therefore see two possible options. The first, more demanding one, would be to restructure the paper using the evaluation runs as the main reference, which in my view would make the manuscript shorter, clearer, and better connected to both the evaluation and projection papers. The second, less extensive option would be to include the evaluation runs in the current comparisons, which would at least help distinguish structural biases from those introduced by the LBCs.

Answer: Thank you for the concrete suggestions on improving the manuscript.

Disentangling the different sources of biases is definitely interesting and important to learn about individual RCMs and the design of RCM ensembles. However, we consider a detailed analysis in that regard beyond the scope of this manuscript. E.g., aspects of this with a focus on the EUR-12 NUKLEUS simulations conducted with ICON and CCLM are discussed in Cusinato et al., 2026 (under review), who investigate the impact of GCM forcing biases on RCM surface fluxes, which may lead to locally reducing temperature biases at the expense of increasing moisture biases. The project NUKLEUS, is an application oriented endeavor, and the manuscript was written with the scope of pointing out how well the ensemble represents the historical climatic conditions over Germany.

Nevertheless, we think that relating the biases more strongly to their potential sources adds value to the manuscript and we plan to incorporate the separation of the different origins of the biases more consistently in the interpretation of our results. In order to do so, we will refer to RCM-specific and forcing-induced biases. In our understanding RCM-specific biases originate from the model assumptions and their implementation, how LBCs are provided to the model and how the models are tuned. The forcing-induced bias originates from potential GWL mismatch and the LBCs, i.e. how a climatic state for a given GWL is represented in the forcing data (e.g., CWTs).

We will shape the discussion of the biases presented with our manuscript based on the assumption that in the evaluation simulations presented in Sieck et al., 2026 there is no GWL mismatch as well as perfect LBCs provided by ERA5 and thus the evaluation simulations exhibit only RCM-specfic biases. Besides the change of the primary references for some of the analyses mentioned above, we plan to extend some of the figures by additionally showing biases averaged over the different forcing datasets for each RCM in order to facilitate the comparison between historical and evaluation simulations.

Overall we consider these suggestions as a good compromise to improve comparability of simulation results with observational data as well as sharpen the discussion of the biases and their potential origins.

Minor comments:

Line 70-71: I would also added that such method still struggle to reproduced the extremes, which is an important issue from those approaches. See https://arxiv.org/abs/2508.15724 for example.

Answer: Thanks for this comment and the additional reference. We will add a sentence on that in the revised version. "**Furthermore, Zhang et. Al (2025), for example, have showed that AI-based models are not able to reasonably represent the extremes, but numerical models do a better job.**"

Section 2.1.2: Missing information on the double nesting process: Spectral nudging, nesting interval for each step, LBC variables for each steps. Also, please show all nested domains

Answer: We will describe potential differences to the nesting and nudging processes described in Sieck et al. (2026) in order to avoid duplication of the information. We will also provide maps of the entire model domains.

Section 2.1.1: I am not convinced that Figure 1 is necessary. The manuscript does not provide sufficient details on how the figure was produced, and its inclusion somewhat disrupts the overall flow of the paper. I would suggest simply explaining the rationale for selecting these GCMs, as is already partly done in the text, and mentioning that their selection was based on an internal analysis.

Answer: We agree that the selection of the GCMs can be shortened here. As we want to keep this quantitative measure to some extent, we will move Figure 1 to the supplemental material and adjust Sect. 2.1.1 accordingly. Instead, sect. 2.1.1 will then be extended by some more insights in the modeling chain from GCM over 12 km to 3 km (e.g., trends, warming levels) referring to your comments below and as also suggested from the other reviewer.

Line 229: "0.05mm" To me this is a rather uncommon threshold. Why this number and why so small? You applied it daily?

Answer: Yes, the quantile-delta-mapping bias correction was applied on daily data. For setting up the methodology, we followed the approach and recommendations by Cannon et al., 2015 (doi.org/10.1175/JCLI-D-14-00754.1), which also use 0.05mm as dry-day threshold. To clarify the source of this value, we will adjust the text accordingly.

Lines 264-268: More details are needed about PSI. For example "largest extent, and longest persistence" how is that computed?

Answer: As lined out in our general answer above we will to strongly cut section 6.2 (limitations of the bias correction approach) and remove results using the PSI and accordingly the methodological explanations.

Section 3: I am somewhat skeptical about comparisons based on a fixed historical period when the RCM/CPM simulations are driven by GCMs, since the level of warming in the GCMs may differ substantially from that observed over the same period (for example 1981–2010 period corresponds to approximately 0.69°C of warming relative to 1850–1900). Therefore, part of the differences identified as biases may actually result from differences in the underlying warming level rather than from biases introduced through the LBCs.

Answer: We understand your concerns. However, the historical 3km simulations are only performed for this fixed period 1961-1990, which was requested by the project partners from the application side during the co-creation process of the NUKLEUS ensemble design. As shown in the Table 1 below, at least one model (MIROC6) deviates from the other two regarding the state of global warming within the historical period compared to the pre-industrial (1881-1910) period, while the other two are comparable and very close to the observations. Yet, regional warming patterns appear to be more relevant for the model biases over central Europe than the level of global warming as is the case for MIROC6, which exhibits lowest global warming between historical and PI and warmest bias over Germany (see Figure 1 in answers to Referee 1, and Table 2 below)

Table 1: *Climatological global mean temperature (in K) in observations and GCMs for pre-industrial (PI; 1881-1910) and historical period (HIST; 1961-1990), and the differences (DIFF) between HIST and PI.*

Data	PI (K)	HIST (K)	DIFF (HIST-PI) (K)
HADCRUT5	286,72	287,14	0,42
ECE	286,88	287,37	0,49
MIR	288,39	288,61	0,22
MPI	287,05	287,51	0,46

Table 2: *Climatological mean temperature (in K) in Germany in observations and GCMs for the historical period (HIST; 1961-1990), and the Bias towards HYRAS for HIST. All data re-gridded to CEU-3.*

Data	HIST (K)	Bias (HIST) (K)
HYRAS	281,42	x
HADCRUT5	281,35	-0,07
ECE	280,76	-0,65
MIR	282,73	1,32
MPI	281,04	-0,38

Line 285-286: This is debatable…

Answer: We sharpened our description here in order to clarify our statements: “For each driving GCM, the large-scale pattern of the biases are very similar in the CCLM and ICON simulations with on average slightly lower values in the ICON simulations (cf. Fig. S10). On the local scale, the biases for ICON simulations exhibits larger variability compared to the CCLM simulations.“

Line 295: Daily temperature extreme > Daily max temperature …

Answer: This paragraph includes statements for both daily maximum and minimum temperatures, so speaking of “temperature extremes” in the introductory sentence, from our perspective, is fine.

Line 309 : Spatial gradients …temperature. > Irrelevant

Answer: We agree. The sentence will be removed.

Line 310: Zonal gradient?

Answer: We will replace this by "**west-to-east gradient**"

Line 313-314: "These biases … unsmoothed topography applied in REMO." So, CCLM and ICON are smoothed? I guess this is written in Sieck et al, but this could also be added in the model description. On the other hand, I am surprise to see that topographic details are low in CCLM and ICON. Why?

Answer: Technically, ICON and CCLM apply a filtering of strong gradients in the subgrid scale orography scheme, e.g., ICON uses a one step smoothing of grid-scale orography to avoid too steep slopes between adjacent grid cells.

Regarding the low topographic details in biases for CCLM and ICON: The fewer topographic details in the bias plots results from relatively low biases in these regions.

Figure 6: y-axis => Monthy accumulated sum. Why using such large temporal scales. Added-value of kilometers simulation would be better shown for daily or hourly time scales. I do not understand this choice.

Answer: In principle we agree that the added value of high-resolution simulations is expected to be more evident on shorter time scales. However, the added value of 3 km vs. 12 km is not within the scope of this manuscript. The overall goal of the entire manuscript is to demonstrate that the 3km NUKLEUS simulations are capable of reasonably representing the climate and natural variability within the historical reference period when using the GCMs as forcing data (instead of ERA5 reanalysis as done in the accompanying study by Sieck et al.). Therefore, the analyses are performed on various spatial and temporal scales. Fig. 6 is part of the section on interannual and interseasonal variability, for which in our opinion monthly values are appropriate. Subsequent sections on frequency distributions and diurnal cycles already deal with shorter time scales.

Line 371-380: Not clear, please reformulate better.

Answer: Agreed, we will rephrase this. More specifically, we will merge the description of the annual cycles shown in the supplemental Fig. S20, i.e. lines 371 to 374, with the description of the annual cycles in the main part (Fig. 7), i.e. lines 357 to 370. For lines 375 to 380 we will rephrase more clearly in the context of discussing the RCM-specific biases and the forcing-induced biases.

Line 386: The rather small biases in summer further tend to be negative … => Not really.

Answer: We agree that this is misleading. We will rephrase this to: "**Throughout the year, biases are predominantly positive. Negative biases mainly occur during summer and exhibit lower in magnitudes compared to the positive biases.**"

Line 401-402: "…here is expectedly larger than in Sieck et al .(2026)." I guess what you means is that the total bias is larger than the structural bias alone? If so, please be more explicit, so the readers doesn't have to reach the other to understand.

Answer: Yes, we consider the biases in Sieck et al. (2026) to originate to a large part from the RCMs. We will point this out more clearly in the revised manuscript.

Line407-410: I wonder then if it would not be useful to add those deviation on figure 9.

Answer: As mentioned in our general answer above we decided to include the TRY wind data for the period 2005-2014 (evaluation period in Sieck et al., 2026) into the wind-related plots (annual/diurnal cycles, pdfs) of the manuscript in order to address a couple of comments made by both reviewers.

Line 420-421: It might be the colors on my computers, but this is not what I see, It seems to hold for the coldest part of the distributions, but not elsewhere…

Answer: We agree and will rephrase this to: Towards the cold end of the distributions (approximately for tas smaller than 0 °C) the three MIR-driven simulations (shown in blueish tones in Fig. 10 and S23) exhibit the highest relative occurrence and also exhibit the minimum temperatures. This is partially also true for the warm end of the distributions, i.e. for tas larger than 17 °C to 20 °C in EF, DUI, and EE, though the most extreme warm values are not predominantly originating from MIR-driven simulations.

*Line 435: I am not fully convinced by this interpretation. A simulation with a 3-km grid spacing still represents quantities averaged over a roughly 3 * 3 km grid box and over the entire atmospheric column. It cannot explicitly resolve the much smaller spatial scales within individual convective cells.*

Answer: In principle, you are right. What was meant here is that RCMs simulate the entire life cycle of convective cells. In contrast, there are lots of convective cell not captured by HYRAS, because they never pass a ground-based station. And even when a convective cell passes a station, this does not necessarily happen at the time of highest intensity. We will rephrase this part to make it clearer.

There is also the question of the observational reference. In a way you used HYRAS, it's become a 2- steps interpolated product (one from the obs for creating the gridded dataset, and another done by the authors for this

analysis), which might have smoothed the extremes. Moreover, given that the observed event reached approximately 353 mm, is a maximum of around 400 mm in MIR-ICON necessarily that surprising? I am not suggesting that your interpretation is incorrect, but I think several other mechanisms could potentially explain the discrepancy. For example, differences in the large-scale conditions provided by the driving GCM could affect the conditioning and development of such an extreme event.

Answer: Thank you. We will include these additional considerations in the manuscript either in the discussion of Fig. 11 or in the overall discussion section.

Line 458: PPS, please do define it somewhere on the methodology section.

Answer: From our point of view, PSS is an established metric and already mentioned in the method section with respective reference. However, we will include an additional explanatory sentence.

Line 537: …" reveal similar median values…" I am sorry, but how it this a relevant. The indices is constructed so the median should be around 10%. The only thing interesting about 16b, it's the comparison about variability. Same comment about line 541 ("this does not mean….).

Answer: We agree and will rephrase to: By design the distributions of the annual hot days TX90p (Figs. 16b and S29b) reveal very similar median values among all NUKLEUS ensemble members, HYRAS, and ERA5-Land. The simulated interannual variability compares well with the observed variability. A slightly larger interannual variability can be found for the MPI-REMO simulations in PAR, DUI, and STU, and the ECE-REMO simulation in EF.

Line 576: I am not sure I am convinced of the utilities of Bias-correcting in the light of my general comment (see above).

Answer: If we interpret your reference to your general comment correctly, you are doubting the value of presenting the results of a standard bias correction procedure within this manuscript instead of focusing the manuscript on the different sources of biases and not the general requirement of bias correcting the data. As lined out in our answer above, the main goal in NUKLEUS application oriented climate information. Thus, especially for deriving climate change signals of (threshold based) climate indices as done in Beier et al. (2026; https://doi.org/10.1175/JCLI-D-25-0613.1), bias correction is indispensable. As the here presented manuscript is intended to provide the foundation for the results presented in Beier et al. (2026), we consider a brief presentation of the (admittedly obvious) benefits and limits of the applied bias correction procedure as necessary and valuable.

Section 5: It this section useful? I fully support comparing the NUKLEUS ensemble with previous ensembles. However, as currently presented, the comparison mainly highlights differences in biases, which is not particularly surprising given the differences between the ensembles. I would suggest briefly summarizing the main differences and then focusing on what we actually learn from the NUKLEUS ensemble. In other words, what new

insight does this ensemble provide that was not already known from previous studies?

Answer: Thank you for questioning the usefulness of this section. In the revised manuscript, we will focus the comparison stronger on the novelties introduced by the NUKLEUS ensemble compared to previous ensembles and cut the comparisons between the ensemble biases, which deserve more detail than in the current version, and which we consider far beyond the scope of this manuscript. Hence, we decided to shorten this comparison and move the content to a paragraph within the general discussion.

Line 606: How I read this line is that HadGEM and MIR are similar? Which does not make sense, maybe reformulate those sentences?

Answer: We agree that the formulation is confusing here. As lined out in our answer above, the detailed bias-specific comparisons of the NUKLEUS ensemble with other ensembles will not be part of the revised version of the manuscript anymore.

Section 6: Despite the fact that I am not convinced that BC is needed here (see general comment above), I also question whether this entire section is necessary. Section 6.1 mainly shows that BC removes the overall bias, which is essentially the expected behavior of a bias-correction method, while Section 6.2 shows that BC can introduce non-negligible artifacts when analyzing individual events. Taken together, I found this section more confusing than informative. In my view, this issue could constitute a separate study on its own.

Answer: As lined out in our answer regarding the bias correction above, we consider a brief evaluation of a bias corrected dataset necessary in the line of presenting the results from the NUKLEUS ensemble, particularly the climate change signals of climate indices presented in Beier et al. (2026). We agree, that the results presented on the limitations of the bias correction approach are expected and do not provide novel scientific insight. We therefore consider moving section 6.2 to the supplemental material and only briefly mention the limitations in the main part.