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.

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.

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.

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.

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

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.

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

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

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.

Line 285-286: This is debatable…

Line 295: Daily temperature extreme > Daily max temperature …

Line 309 : Spatial gradients ...temperature. > Irrelevant

Line 310: Zonal 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?

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.

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

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

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.

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

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…

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. 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.

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

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….).

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

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?

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

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.