

NUKLEUS – A First Kilometre Scale Multi-model Climate Ensemble for Germany: Evaluation

Referee #1

In the following, GMD review criteria are individually addressed by the reviewer, item by item (see https://www.geoscientific-model-development.net/peer_review/review_criteria.html); some are just comments to the editor, others are also meant as suggestions to the authors for improvements or clarifications:

"1. Does the paper address relevant scientific modelling questions within the scope of GMD? Does the paper present a model, advances in modelling science, or a modelling protocol that is suitable for addressing relevant scientific questions within the scope of EGU?"

The paper is suitable in form and content for a publication with GMD.

The authors present a new simulation experiment, a multi-model regional climate change ensemble with state-of-the-art RCMs at convection permitting resolution. Such simulations are still relatively rare, especially when done in a concerted manner.

There is much proven added value when going from convection-parametrized to convection-permitting resolutions at km-scale grid spacings for the reproduction of small scale heterogeneity, surface atmosphere interactions, convective processes, reproduction of dynamical features, etc.

The data produced with the experiment design or modelling protocol supports an advancement of regional climate modelling science as this is the first km-scale ensemble for the region of Germany. The data generated may be used for further processes and VIACS studies.

The experiment thereby produces evaluated (not optimized) model configuration and setup information, gives evidence of model fidelity, provides a simulation protocol synchronized with relevant stakeholders, and certainly also uncovered and helped resolve many HPC-related technical obstacles.

The experiment design as well as the analysis follows best practice approaches without being overly innovative (one-way double nest, reanalysis downscaling, NWP model configurations). The meticulously done evaluation, with suitable reference data,

systematically pinpoints model deficits for ensuing model improvement or data preparation (bias adjustments).

As convection-permitting simulations seem to become a new standard, experiments such as NUKLEUS help to gain experience in all aspects of running and maintaining such simulations (runtimes, configurations, data volumes, reference data, model behaviour, deficits in inputs and parametrisations) and thereby promote such simulations.

The paper does not make an attempt for an in-depth explanation of biases. This may be a weakpoint but can be considered valid, as the focus is strictly on evaluation. Likewise no computational aspects or "lessons learned" are addressed. This in essence manifests itself with simulation information (i.e., how were the simulations run, configurations, input data, see review item 6).

Answer: Indeed, we did not dive too deep into biases, which seemed as pointed out too much with our focus on evaluation. Wherever we discovered known issues of convection-permitting simulations or of single models, we tried to refer to them. Considering the "lessons learned" we are happy to share more information (see also answers to items #5 and #6).

"2. Does the paper present novel concepts, ideas, tools, or data?"

The authors follow a well proven experiment design, starting off with evaluation run as a basis for a follow up climate change scenario simulation.

Model systems are state-of-the-art; many users will most likely be interested in the performance of the ICON-CLM, which can be considered the next generation model over COSMO and REMO.

The generated data is novel, see review item 1.

The specific focus on and evaluation for six pilot regions across the model domain, based on a stakeholder co-design approach, can be considered novel (here more information would be interesting on the process, L.100ff, but certainly beyond the scope of the paper). In the evaluation, the specific characteristics of the pilot regions are addressed in the interpretation of the biases.

Aside from only assessing thermal and hygric quantities, 10m wind speed and global radiation seem clearly tailored to renewable energy, using km-scale information for wind energy potential analyses.

It can be considered a novel concept in regional climate modelling to run coordinated km-scale climate change ensembles. NUKLEUS facilitates that such experiments become the new normal in regional climate modelling.

Answer: We could add one more sentence to the stakeholder process. Our suggestion would be: "The focus of the workshops was on suitable input for the climate adaptation tools developed in the pilot regions."

"3. Does the paper represent a sufficiently substantial advance in modelling science?"

Yes, given the combined configuration refinements or tests, newly produced datasets, and the thorough model evaluation and testing.

Modelling science could be more advanced if the authors provide more information on computational and big data aspects, as this is often a hindrance for such model runs, and lessons learned from the setup and configuration.

Answer: We are happy to add more details on the configuration and computational aspects. We will add them to several sections. Details for the configuration will go to the supplement (see answer to item #6). Lessons learned from the computational aspects will be added to the conclusions, which we plan to rework and add more information on common features of these simulations with earlier work (see answer to item #2 of referee #2).

As mentioned above, the experiment is novel. Within the GCM realm, I am struggling whether this is to be considered a "model experiment description paper", as the information on the exact experiment details may be a bit vague (see review item 6), also I would not consider this a generic benchmarking experiment; to me this appears more a "model evaluation paper" that provides "thorough performance and behavior evaluations of previously published models".

The discussion remains in places vague when it comes to addressing root causes of the biases and model behaviour and concludes that despite the substantial improvements we see, bias adjustment is still needed for many VIACS applications. This is OK, as the study focusses strictly on evaluation and a thorough assessment is beyond the scope, but perhaps make this very explicit to not raise false expectations with the readers.

Answer: We suggest adding a sentence to the abstract in line 5: "Although partly discussed, an in-depth discussion of individual model biases is beyond the scope of this paper."

"4. Are the methods and assumptions valid and clearly outlined?"

The overall experiment design is clearly explained, more details would help in reproducibility (see other review items). The dynamical downscaling, with a one-way double nest, is an established procedure. The evaluation follows in terms of variables considered and diagnostics and statistical measures common practice, incl. also the ETCCDI indices. I was surprised that you did not show any Taylor diagrams? The reference data are well chosen.

In addition to the evaluation for the six pilot regions, more information on all of Germany would have been interesting. The reviewer assumes that the pilot regions are considered representative enough. Why do you show the EF and TSA regions in the main text? The STU domain would have had a similar number of meteorological stations as EF and be similar in size.

Answer: Concerning the choice of domains for the analysis in the main part, we wanted to have representative regions for different climate types in Germany. EF being more maritime influenced and a representative of the northern lowlands and TSA representing a more continental and typical German low mountain-range region. Especially, in case of the low mountain range region the choice could have been STU, too. In the end, TSA seemed to be a more geographically balanced choice. We suggest changing the sentence in line 301 to: "The main part of the paper presents figures only for two of the six pilot regions, namely EF representing the German lowlands and TSA as representative for a typical German low mountain-range region. Results from other regions will be discussed, if worth mentioning."

As the paper shall define a specific simulation protocol, serve as a benchmarking, why was this specific timespan chosen?

Answer: The timespan 2005-2014 was chosen to provide the best overlap with the various observational datasets used for the analysis.

The reader inevitably will ask why the 12km intermediate nest, for the well proven and widely used EURO-CORDEX domain, is not also covered to show the added value on-the-fly, see also a comment with the review item 5 on this. So perhaps mention the added value paper also earlier.

Answer: We will add the information on the other papers already in the Introduction (see answer to item #8).

"5. Are the results sufficient to support the interpretations and conclusions?"

In principle yes. The rigorous evaluation uncovers systematic biases for some of the most relevant stakeholder variables. Overall there is a good agreement with observations. The NUKLEUS experiment certainly also provides "a substantial advancement in high-resolution climate information for Germany" (L.517f). The conclusions on model improvement mainly rely on a comparison with evaluation studies from literature.

Whether substantial added value is provided by the km-scale runs is to be expected, but remains from the analysis IMHO unclear, so "...offering improved representation of convective processes, regional climate features, and local extremes compared to coarser regional or global models." (L.518f) would need to be rephrased; the 12km intermediate nest vs 3km comparison is done in a separate companion paper.

Answer: We agree that with this paper it is hard to argue about improvements. We would rephrase the sentence from "improved" to "more detailed".

Although I think the statement as such is valid, that "Overall, the ensemble demonstrates strong potential as a foundation for actionable climate information, especially in the context of regional climate adaptation efforts within the RegIKlim program." (L.522ff), because the stakeholder requirements wrt quality criteria for their applications (a core motivation for the production of the ensemble) are not mentioned, there is no proof the data is actually fit for purpose, however, the authors cite Pinto et al. (2026, <https://doi.org/10.1002/joc.70304>) in the context of climate indices. Maybe rephrase this and/or further elaborate.

Answer: Defining quality criteria for the applications in the model regions is hard since they are very diverse. Two studies by Laux et al. (2025, <https://doi.org/10.1016/j.jhydrol.2025.133607>) and Wagner et al. (2026, <https://doi.org/10.5194/hess-30-3597-2026>) using NUKLEUS data have been published so far. Wagner et al. (2026) show that there are benefits for hydrological modelling in terms of temperature, radiation and wind speed, but that the deficiencies especially in precipitation amounts does not result in an overall added value compared to coarser resolved data. Laux et al. (2025) took a step further and used bias correction for the precipitation amount. They show that the hourly data at 3km provided by NUKLEUS allowed them to estimate future flood risks at very high resolution. This highlights our statement that we still see room for improvement in our models despite making progress. We suggest adding the following sentence to line 522 before the cited statement: "This is also highlighted by two studies undertaken in the pilot regions by Laux et al. (2025) and Wagner et al. (2026) who find benefits using the data from convection-permitting simulations done in NUKLEUS but also show that bias correction is still needed for certain applications."

From the evaluation, it seems that ICON-CLM, not surprisingly perhaps, may be the best performing model, did this have implications for the further experiment design of NUKLEUS?

Answer: It did not change the experiment design of NUKLEUS but influenced the design of companion projects such as UDAG (https://www.dwd.de/DE/forschung/projekte/udag/udag_node.html). In NUKLEUS we wanted to take the opportunity to test the legacy models against the next-generation model (ICON-CLM). NUKLEUS is the first project where ICON-CLM was used following the full climate change simulation protocol in regional climate modeling. In the follow up project UDAG, only ICON simulations were conducted and the lessons learned from NUKLEUS were used in the experiment design, e.g., the feasibility of transient simulations over 150 years. There is a paper on UDAG by Singh et al. in review.

"6. Is the description sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)?"

As this is a "model experiment description paper", the simulation experiments needs to be reproducible. At the time of writing this review, the discussion is open on how model source codes, input data, results datasets may be made publicly available to ensure reproducibility (<https://doi.org/10.5194/egusphere-2026-1024-CEC1> and <https://doi.org/10.5194/egusphere-2026-1024-AC1>). A similar discussion is ongoing with the companion paper on the historical simulations by Braun et al. (<https://doi.org/10.5194/egusphere-2026-2517-CEC1>). At this point the information provided in "Code and data availability" may be considered not sufficient to meet the requirements in as defined in https://www.geoscientific-model-development.net/policies/code_and_data_policy.html.

Aside from formal GMD requirements, a reproducible experiment is highly desirable to help the community to pick up this type of experiment and get started more quickly with their own km-scale workflows.

As the reviewer understands, there are several aspects and possible solution: (i) <https://www-regiklim.dkrz.de> will provide FAIR open access to the NUKLEUS data, right now data is free but access is restricted; in any case a Zenodo-based table (with DOI) may provide a listing of URIs to make the model results data files used in the study findable; (ii) the specific model versions may be made available as software packages (with DOI) via Zenodo; (iii) the input data to the models (static and external parameter fields, etc.) may also be shared through Zenodo with a DOI; (iv) the model configurations (i.e. parametrisation settings, etc.) as well as the scripts for pre- and postprocessing are packaged in case of ICON-CLM and COSMO-CLM via the SPICE and STARTER-Package integrated workflow engines, available through Zenodo; perhaps this has to be more clear in the main text as well as in the "Code and data availability" section; are the REMO configurations available somewhere public?; (v) evaluation analysis tools are at least to some degree part of FREVA package, FOSS available through <https://zenodo.org/records/1325149> and <https://github.com/freva-org/freva-legacy>; (vi) most reference data (see also the discussion for Braun et al. companion paper, <https://doi.org/10.5194/egusphere-2026-2517-CEC1>) is available as FAIR, free, open access research data from their respective repositories.

How to deal with this situation of especially (i), the reviewer leaves open to the editor.

Answer: We were dealing with this issue and published the requested data and code in suitable repositories (see answer to chief editor, <https://doi.org/10.5194/egusphere-2026-1024-AC2>).

Maybe structure the information in the Code and Data Availability section more clearly, so a reader not familiar with COSMO-CLM or ICON-CLM procedures around Starter Package and SPICE can easily understand.

Answer: We will rewrite this section adding all information also on the data and code sources mentioned above.

Despite the principle availability of the configurations through the workflow engines for COSMO-CLM and ICON-CLM, perhaps add tabulated information to the supplement, which highlight the most relevant non-default configuration settings, also related to the km-scale resolution. Despite the description in Section 2.1 (e.g., L.150ff), the exact configurations are referred to being close to the NWP (L.150, L.190), but as a foundation paper for the follow-up companion papers, it might be very helpful to provide more information.

Table 1: List of ICON namelist parameters deviating from the default

Namelist	Parameter	Default	ICON 3 km
run_nml	ltestcase	.TRUE.	.false.
run_nml	iforcing	0	3
run_nml	ltransport	.FALSE.	.true.
run_nml	ntracer	0	5
run_nml	num_lev	31	60
run_nml	dtime	0.	100
run_nml	timers_level	1	10
run_nml	msg_level	10	0
run_nml	output	default	nml
run_nml	check_uuid_gracefully	.FALSE.	.TRUE.

nwp_phy_nml	mu_rain	0.0	0.5
nwp_phy_nml	rain_n0_factor	1.0	0.1
nwp_phy_nml	inwp_radiation	1	4
nwp_phy_nml	latm_above_top	.FALSE.	.true.
nwp_phy_nml	ldetrain_conv_prec	.FALSE.	.true.
nwp_phy_nml	efdt_min_raylfric	10800.	7200.
nwp_phy_nml	icapdcycl	0	3
nwp_phy_nml	icpl_aero_conv	0	1
nwp_phy_nml	icpl_aero_gscp	0	1
nwp_phy_nml	dt_rad	1800.	900
nwp_phy_nml	dt_conv	600.	300
nwp_phy_nml	dt_sso	1200.	600
nwp_phy_nml	dt_gwd	1200.	600
radiation_nml	ecrad_data_path	.	data
radiation_nml	ghg_filename	bc_greenhouse_gases.nc	greenhouse_ssp370.nc (eigener Pfad)
radiation_nml	irad_co2	2	4
radiation_nml	irad_ch4	3	4
radiation_nml	irad_n2o	3	4
radiation_nml	irad_cfc11	2	4
radiation_nml	irad_cfc12	2	4
radiation_nml	irad_o3	0	79
radiation_nml	irad_aero	2	6
radiation_nml	albedo_type	1	2
radiation_nml	direct_albedo_water	4	3
radiation_nml	albedo_whitecap	0	1
nwp_tuning_nml	tune_gfrcrit	0.4	0.425
nwp_tuning_nml	tune_dust_abs	0.0	1.0
nwp_tuning_nml	tune_zvz0i	1.25	0.85
nwp_tuning_nml	tune_box_liq_asy	3.0	3.25
nwp_tuning_nml	tune_gfluxlaun	2.50e-3	3.75e-3
nwp_tuning_nml	tune_rcucov	0.05	0.075
nwp_tuning_nml	tune_rhebc_land	0.75	0.825
nwp_tuning_nml	tune_gust_factor	8.0	7.0
nonhydrostatic_nml	vwind_offctr	0.15	0.2
nonhydrostatic_nml	damp_height	45000.	10500.
nonhydrostatic_nml	rayleigh_coeff	0.05	1.0
nonhydrostatic_nml	divdamp_order	4	24
nonhydrostatic_nml	divdamp_fac	0.0025	0.004
nonhydrostatic_nml	divdamp_type	3	32
nonhydrostatic_nml	thslp_zdiffu	0.025	0.02
nonhydrostatic_nml	thhgt_dzdiffu	200.	125.
nonhydrostatic_nml	hbot_qvsubstep	22500.	16000.

Answer: We would add Table 1 on the ICON settings and a similar table for COSMO-CLM to the supplement.

"7. Do the authors give proper credit to related work and clearly indicate their own new/original contribution?"

- In principle yes, especially from L.77ff onwards similar other experiments, like the CORDEX FPSCONV, are mentioned. Perhaps expand this overview by other similar modelling activities to NUKLEUS in Europe, like the simulation experiments with HCLM at 3km over Fenno-Scandinavia (e.g., Lind et al., 2022, <https://doi.org/10.1007/s00382-022-06589-3>) or the 2.2 km UKCP18-CPM UK ensemble (Kendon et al., 2020, <https://doi.org/10.1175/JCLI-D-20-0089.1>). Also, previously km-scale time-slice climate

projections and evaluation runs for Germany or sub-regions exist, though no large ensembles, perhaps also mention those in this context, like Fosser et al. (2016, <https://doi.org/10.1007/s00382-016-3186-4>) or Knist et al. (2020, <https://doi.org/10.1007/s00382-018-4147-x>) and others.

Answer: We are happy to expand the list of related work with the given papers. We will change the paragraph starting in line 35 in the following way (changes are marked in bold-face):

As adaptation is mostly happening at regional to local scales, the demand for very high resolution climate information in order to support decision making is growing. To date, the data provided by global climate model (GCM) simulations is insufficiently resolved to satisfy this demand. However, recent advances in computer power and model development made regional kilometre-scale climate simulations feasible (**Kendon et al., 2020**; Kendon et al., 2021; Lucas-Picher et al., 2021; **Lind et al., 2022**). Besides the pure technical feasibility of these simulations, they come with several advantages compared to coarser climate simulations in terms of the representation of physical processes (e.g., Prein et al., 2015; Coppola et al., 2018; Soares et al., 2024). One of the processes that has important societal impacts and can be better represented in detail is convective precipitation. In kilometre-scale climate simulations, convection is explicitly resolved in most models, which often leads to a much better representation of small-scale high-intensity precipitation events compared to models with coarser resolution. Recently, many studies have shown the often called “added value” of convection-permitting climate simulations compared to coarser resolved models. Large benefits are expected for the intensity distribution and diurnal cycle of precipitation (Lucas-Picher et al., 2021; Fowler et al., 2021; Hundhausen et al., 2024) due to the explicit treatment of convection. An added value has also been demonstrated for temperature (Soares et al., 2024; Hundhausen et al., 2023), near-surface wind speed (Molina et al., 2024), and for user-relevant climate indices (Pinto et al., 2026). **First studies to perform convection-permitting climate simulations for Germany date back to Fosser et al. (2017) covering southwestern Germany and Knist et al. (2020) covering a central European domain including entire Germany.**

"8. Does the title clearly reflect the contents of the paper? The model name and number should be included in papers that deal with only one model."

- Yes. Given there is a companion paper ("NUKLEUS – A first kilometer-scale convection-permitting multi-model climate ensemble for Germany: Characteristics of the historical simulations 1961–1990", see above), perhaps indicate this earlier on than just in the outlook. Although the ensemble is available for all of Germany, the title implies that the evaluation exercise covers all of Germany while in fact only the pilot regions are considered in detail.

Answer: We will add a sentence in the introduction (line 75) and mention the companion paper by Braun et al. (2026) and the published work on the future scenarios by Beier et al. (2026): "A detailed discussion of the NUKLEUS historical simulations can be found in Braun et al. (2026) whereas the future climate change runs are described in Beier et al. (2026)."

"9. Does the abstract provide a concise and complete summary?"

- Yes.

Answer: OK

"10. Is the overall presentation well structured and clear?"

- The manuscript follows a very clear, easy-to-follow structure. For example, the results sections 3.1 and 3.2 on temperature and precipitation are identically structured (spatial bias maps, annual cycles, bias plots, 2d bias plots (with MB, MAD), freq distributions (with PSS + MAQD), diurnal cycles). Few things might be improved as follows:

- Perhaps in L.86ff call this part simulation setting or framework, as this gives the overall experiment design.
- In L.130ff, there is a certain imbalance in model description sections between ICON-CLM, and COSMO-CLM and REMO, which are much shorter. This may be due to the fact they models are better known and ICON will be likely superseding the models, but as all three RCMs are used in the NUKLEUS ensemble, perhaps the the model descriptions can be somewhat similar in structure and extend; see also review item 6.

Answer: Referee #2 made a suggestion on the title of the section as well, and we will follow referee #2 here. The main reason for keeping information on REMO and COSMO-CLM shorter was already mentioned. We are happy to rework the sections on the model descriptions by making them more coherent and add some details on the older models were appropriate.

"11. Is the language fluent and precise?"

- The manuscript reads well and fluent. Language is clear and precise. There's only a few specific remarks at the end of this review.

Answer: Thank you! We will of course make the suggested changes given in the specific remarks.

"12. Are mathematical formulae, symbols, abbreviations, and units correctly defined and used?"

- Yes. But perhaps provide more complete variable names, L.74ff, like total precipitation, 2m air temperature, global radiation, 10m wind speed.

Answer: The text will be changed accordingly: "The focus lies on the evaluation of basic meteorological variables in convection-permitting multi-model (CPM) simulations against observational datasets, namely near-surface air temperature, total precipitation, global radiation, and 10m wind speed."

"13. Should any parts of the paper (text, formulae, figures, tables) be clarified, reduced, combined, or eliminated?"

- Fig.1: Expand the figure to each side to show subset maps of the pilot regions incl., e.g, the DWD stations and land use and topographic information (isolines) perhaps; you argue that spatial heterogeneity and small scale features may be better resolved, this can be shown here. Also the Pilot regions are crucial for the evaluation, e.g., urbanisation may be important. Also add some information on the size, I think this is

missing. Maybe this is described in another NUKLEUS paper, but as this paper deals with the "experiment description", this information is valuable for the reader.

Answer: We added the stations to Fig.1 (see below) and changed the caption (bold-face part) accordingly. To get a better inside into the regions we suggest adding a Figure as S1 to the supplement with all regions as magnified sub-maps and isolines to better see the structure of the specific regions. Including land use does not make too much sense in our view at this point, because the land use data sets used and how land use is treated inside the model is very different. This is one of the main issues in current downscaling approaches, because there is a lack of input data at such high resolution, especially when it comes to the historical and future scenario periods.

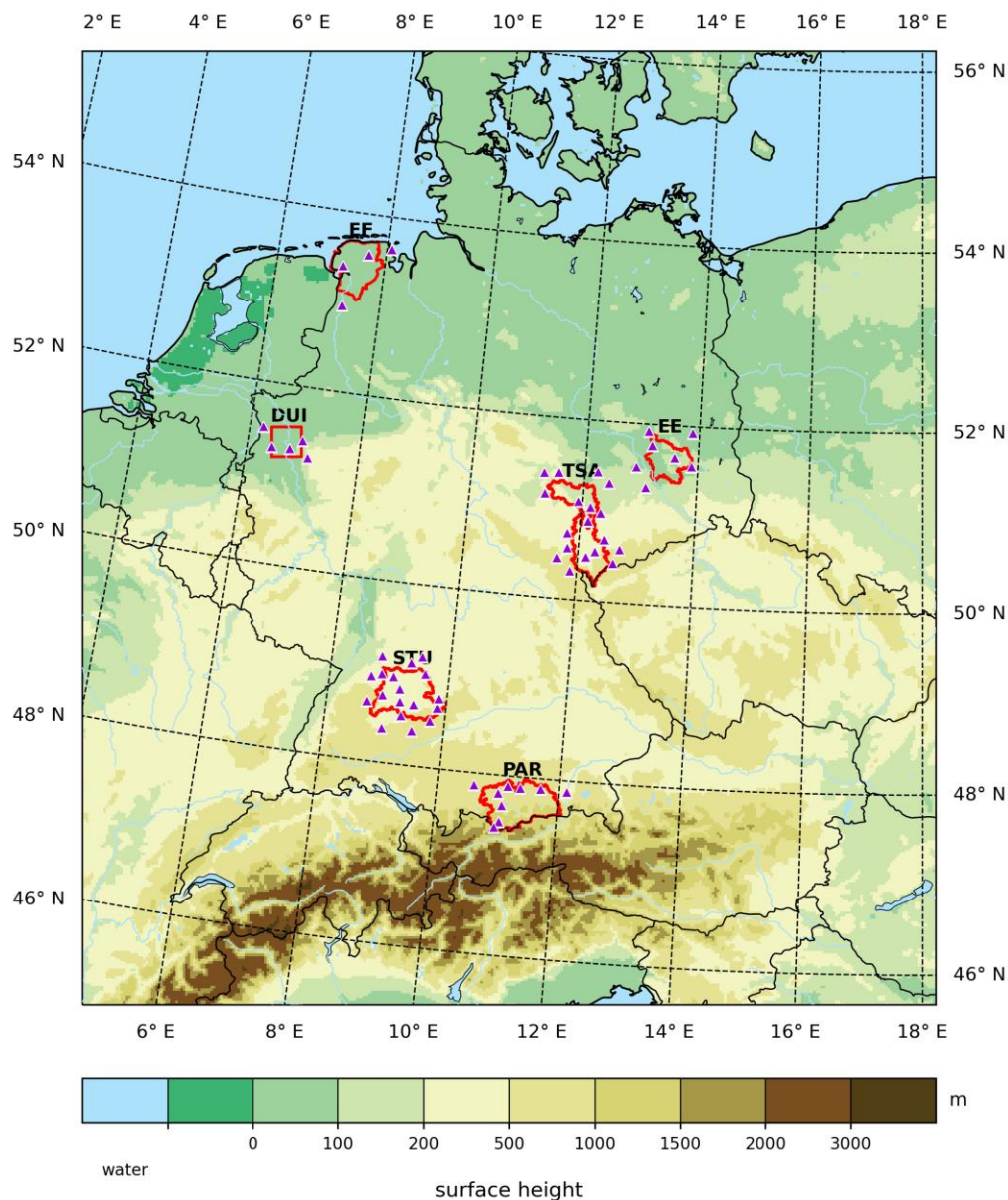


Figure 1: Central European domain (CEU-3) with selected pilot regions used for the statistical analysis of simulation results: coastal areas of East Frisia (EF), extended city areas of Duisburg (DUI) and Stuttgart (STU), Pre-Alpine region in South Bavaria (PAR), a county area in the border triangle of the federal states of Thuringia, Saxony and Saxony-Anhalt (TSA) and of the county Elbe-Elster (EE) in Southeast Germany. The locations of observational stations (purple triangles) are added for later use. Coastlines and national borders are shown for geographical reference.

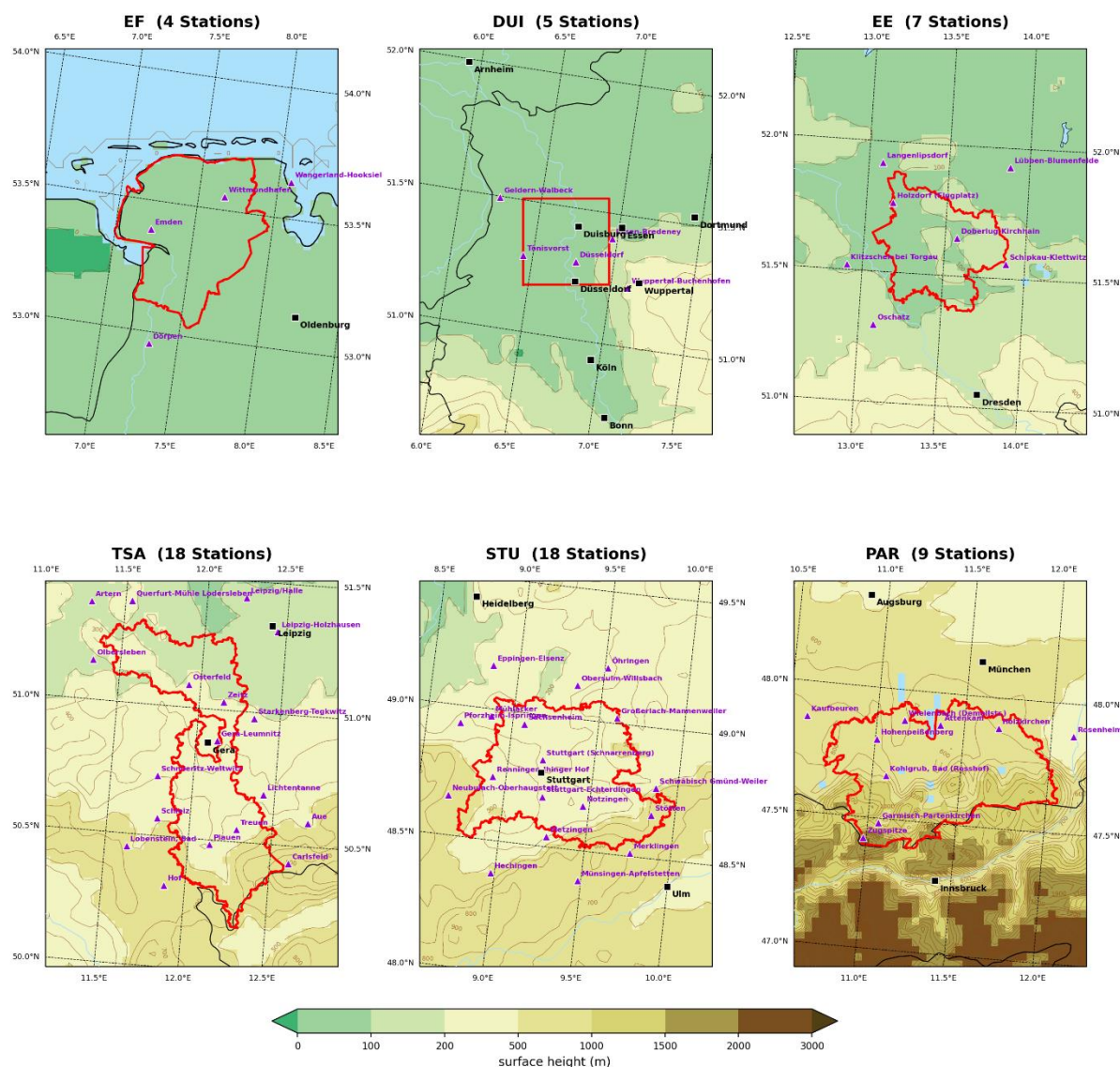
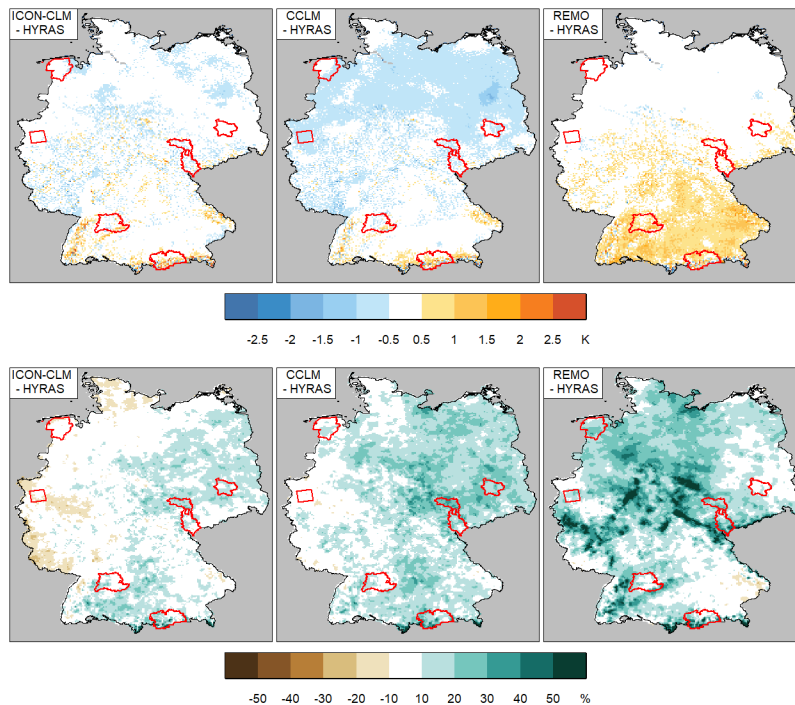


Figure S 1: Detailed orography (colours and contour lines) for the pilot regions (red outlines), including the locations and names of observational stations (purple triangles). Coastlines, national borders, and major cities are shown for geographical reference.

- Fig.2 and Fig.11: Overplot the six pilot region boundaries for orientation.
- Answer: Done accordingly (see below).



- As the paper follows a rigorous structure, see review item 10, this results in many figures. To me some figures might be combined as they are linked. E.g. consider combining temperature analysis Fig.3+4, Fig.5+6, Fig.7+8, Fig.9+10; precipitation analysis: Fig.12+13, Fig.14+15, Fig.16+17; radiation and wind speed perhaps also accordingly (Figs.18-21).

Answer: We see the point of combining some of the figures, but we would like to keep them separate, because the nature of the figures are different (line plots vs. tables) and in some cases like Fig.9+10 it would end up in odd shaped figures.

"14. Are the number and quality of references appropriate?"

- In principle yes, there are a number of statements in the text, which may be backed up by additional references, though, e.g. L.35f.

Answer: We will add the mentioned publications (see answer to 7.)

"15. Is the amount and quality of supplementary material appropriate? For model description papers, authors are strongly encouraged to submit supplementary material containing the model code and a user manual. For development, technical, and benchmarking papers, the submission of code to perform calculations described in the text is strongly encouraged."

- Supplementary information is highly important as the text contains only information on two pilot regions out of six. So the supplementary material is a logic extension / addons to the plots from the main text. Perhaps consider reordering the supplementary figures according

to the sequence in the main text, move S4 to the end. See review item 6, perhaps it would be good to have tables with model parameterisation listings.

Answer: Thank you for the suggestion. We will make the recommended changes.

SPECIFIC COMMENTS, The are on various parts of the manuscript, which do not fir anywhere else:

- *L.25ff: Perhaps also mention the motivate that Europe experienceing accelerated strong climate change and also refer to governmental programmes in which NUKLEUS was run, in addition to the longer explanation in "Method and Data".*
- Answer: We will add a sentence to line 28 "This is particularly important in Europe as the continent warms around three times faster compared to the global mean (Dong and Sutton, 2025)."
- *L.29f: Urban planning should be in a separate sentence, the context, especially through "particularly", with the hydrometeorological extremes sounds misleading.*
- Answer: The sentence will be changed to "for example" instead of "particularly" (see also answer to one of the comments by referee #2)
- *L.49ff: In this context the computational cost vs domain size vs experiment simulation length may be addressed more in detail as this is an important aspect, cite, e.g., papers such as Schär et al. (2020, <https://doi.org/10.1175/BAMS-D-18-0167.1>).*
- Answer: Thank you for pointing this out. We are happy to add the suggested publication and rework the paragraph (see item below).
- *L.50ff: Perhaps reorder this part of the paragraph to distinguish the continental and global short runs vs the longer smaller domain experiments, still often for time slices only. There are all in their own right frontier simulations.*
- Answer: We will do this accordingly.
- *L.52: Provide some examples.*
- Answer: We will add examples such as Knist et al. (2020), Coppola et al. (2018) and Kendon et al. (2020).
- *L.56ff: RCM emulators -> advote a sepearte section, more lit, his is a bit too weak, this is an old text*
- Answer: We agree that this field is evolving very fast and needs more emphasis. We will dedicate an extra paragraph to the emulators and add new papers such as Kendon et al. 2025 (<https://doi.org/10.1175/BAMS-D-24-0114.1>) to the text.
- *L.66: "pioneering" refers here to the regional model runs, maybe rephrase*
- Answer: We will add "regional" to the sentence.
- *L.71: In the context above, what is a "mini-ensemble"*
- Answer: This was an internal term used in EUCP and should not be used here. We suggest rephrasing this to "small ensembles of at least two RCMs".
- *L.76f: Perhaps identify variables under investigation more clearly, like total precpitation, near surface air temperature, etc., although it certainly is intuitive*
- Answer: We will adjust the sentence (see answer to item #12).
- *L.115: Give examples for the application of the models.*

- Answer: We will rework this sentence, because it is not precise. We will add examples on REMO and COSMO-CLM and highlight the first use of ICON in longer term climate simulations.
- L.254: *Put station data in separate paragraph to emphasize this data product.*
- Answer: We will expand information on this product and dedicate an extra paragraph.

MINOR (TECHNICAL) CORRECTIONS

Answer: Thank you for catching these mistakes. We will correct the manuscript accordingly.

- L.290, sentence: "...is slightly from..."
- L.293, typo: "gird"
- Tab.1, caption: TXin90, TNin90 not actually listed in table
- L.296, grammar: "present"
- L.297ff, remove sentence, superfluous: "The latter are..."
- Fig.6, caption, replace: "meaning of ..." -> "See Fig.4 for cell colour coding."
- L.352, references: refer to the Supplement Figs, check that all supplementary Figs are referenced in the main text
- L.483, typo: dates -> data
- L.516, sentence: "...will be the focus..."

Referee #2

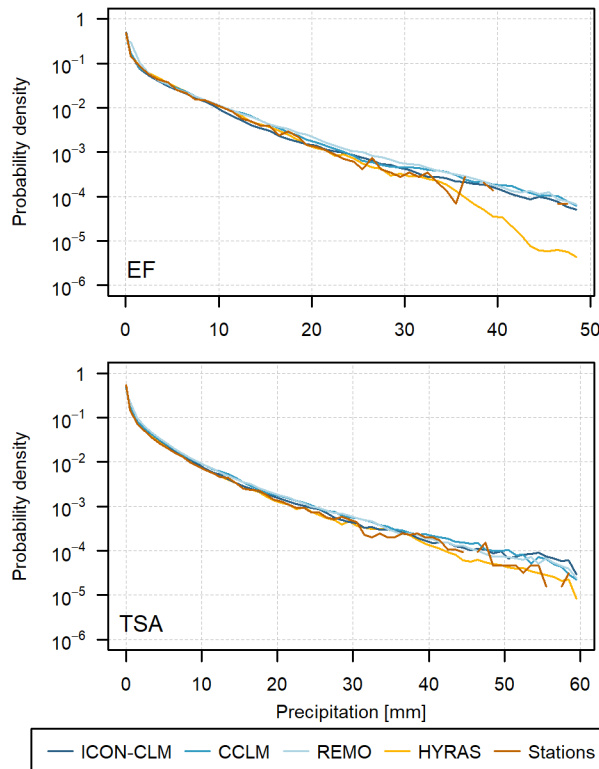
MINOR ISSUES

- *Observational uncertainty: Some coverage of observational uncertainty (i.e. the use of multiple gridded reference datasets where available, or the discussion/explicit consideration of precipitation undercatch in the observation-based products; for instance: is the radar-based RADKLIM product affected by undercatch through assimilation of station observations?) would further improve the significance of the paper.*
- Answer: Thank you for pointing this out. We will add information on observational uncertainty from other sources such as Hammoudeh et al. (2025) (<https://doi.org/10.3389/feart.2025.1548557>) to the discussion.
- *Link to evaluation of further CPM evaluations: Several studies on the evaluation of convection-permitting model ensembles already exist in the available literature. Though these cover different regions in most cases, providing some relation to previous findings would be extremely helpful I believe. Are the biases found in the present work comparable to previous findings?*
- Answer: A similar point has been raised by referee #1 and we are happy to add some more publications to the introduction. In addition, we will add some sentences in the discussion to mention common biases with earlier studies.
- *Comparison to the performance of the intermediate 12km simulations: A comparison of the 3km to the intermediate 12km ensemble would be highly valuable in terms of an added value analysis. As far as I understand, this aspect will be covered by an accompanying publication. One way or the other, this comparison would be extremely relevant and helpful.*
- Answer: As mentioned, we are discussing the added value in a companion paper. Here the focus is on the performance in small regions. For our stakeholders the fact that there is data at kilometer-scale has a value on its own because they are often not able to work with coarser input data.
- *Relevance for climate change analyses: The present manuscript falls somehow short of the discussion and interpretation of the found model performance in relation to the analysis of the GCM-driven ensemble (the latter probably being the main application of the NUKLEAUS ensemble). What are the implications of the findings of the present work for further climate change analyses? The necessity of bias-adjustment is mentioned, and a few points such as multi-model ensemble analysis are listed in the very last paragraph, but only very briefly. Could the authors try to strengthen their manuscript in this respect?*
- Answer: We suggest adding a couple of examples where the NUKLEUS data was used in studies by the pilot regions (see answer to item 7 of referee #1) to strengthen this point.
- *Lines 27-28: You might want to add the present, extremely hot and dry summer of 2026.*

- Answer: we will expand on Europe warming faster and mention the exceptional summer of 2026 (see answer to a comment on line 25 by referee #1) and change the following sentence to mention summer 2026 (changes are in bold-face): “In Germany, events such as **the extreme summer of 2026 (Ionita and Nagavciuc, 2026)**, the extreme drought of 2018 (Rousi et al., 2023), and the catastrophic Ahr Valley flood (Ludwig et al., 2023; Mohr et al., 2023) have intensified the urgency for action, **for example** in the context of urban planning (Fünfgeld et al., 2026).”
- Lines 29-30: “particularly” à *Urban planning is certainly one sector heavily affected, but many further of similar importance exist. Either change to “for example” or list some more sectors.*
- Answer: We will change the sentence as suggested (see above).
- Line 53: *You might want to cite Leutwyler et al 2017 (JGR Atmosphere) in addition here. To my knowledge, they presented the first European scale convection-permitting model setup.*
- Answer: We will add this publication to the list in the sentence.
- Line 54: *Several more recent references would be available here (such as Prein et al. 2026 GMD).*
- Answer: Yes, there are a number of more recent papers on this. Besides Prein et al. (2026) we also add Hohenegger et al. (2023, <https://doi.org/10.5194/gmd-16-779-2023>) to the list of example studies.
- Line 54: *I suggest rephrasing to “these models can CURRENTLY only simulate a couple of months to a few years” (there are strong indications that longer simulations will soon be available, or are already, see the DESTINE climate twins, for instance).*
- Answer: We will change the sentence accordingly.
- Line 85: *I suggest using “METHODS and Data” as title.*
- Answer: We will change the title accordingly.
- Line 100: *“To meet the user’s demands”.*
- Answer: We will change the sentence accordingly.
- Line 104: *A reference to Fig. 1 could be added here.*
- Answer: We will add the figure reference accordingly.
- Line 111-112: *I would use a more neutral formulation here. The analysis should probably present an objective evaluation of the ensembles, whether or not this increases the user’s confidence and the “quality of the provided data” is still to be seen.*
- Line 134: *“PHYSICS package”.*
- Answer: We will correct this accordingly.
- Line 160: *Many readers might not be familiar with the term “TKE”, I suggest using the full name here.*
- Answer: We will use the full name here.
- Lines 152-153: *“greenhouse gas CONCENTRATIONS”.*
- Answer: We will correct this accordingly.
- Line 185: *I suggest specifying that this statement refers to ICON-CLM only: “The ERA5 driven ICON-CLM evaluation runs...”.*
- Answer: We will add the information about ICON.
- Lines 186-187: *I don’t understand why this can be a reason. The ERA5 soil moisture index could still be out of balance with the ICON-CLM soil (different resolution, possibly different soil physics, different soil layering etc.).*

- Answer: The sentence “The ERA5 driven evaluation runs did not use any spin-up procedure for the soil, in contrast to the COSMO-CLM and REMO simulations, as the critical soil moisture index, required for the soil initialization, is directly provided by the reanalysis data.” was describing the procedure for the 12 km simulations and is taken out as we concentrate here on the 3 km runs.
- Line 207: “20 PIXELS”.
- Answer: We will change this as suggested.
- Line 228: Before (Line 162), the FLake model was referred to as the “Flake” model (L not capitalized). Please use just one version.
- Answer: We will use “FLake” in the manuscript and correct occurrences of Flake.
- Line 254: 0.05° and 30 minutes probably refer to the spatial resolution of the data. If so, please indicate this.
- Answer: It refers to space and time. We will change the sentence to: The high-resolution data with 0.05 ° **in space** and 30 minutes **in time** were interpolated to the models ~~resolution in time and space~~ spatial and temporal scales.
- Line 262: “arithmetically averaged” à Are the different lengths of the months taken into account?
- Answer: Yes, the different lengths were accounted for.
- Line 273: You probably mean “optimum values near 1”?
- Answer: The sentence is a bit confusing. We suggest the following change: “For quantification, we calculate the mean absolute hourly difference (MAHD) for radiation, the ratio in the diurnal temperature range (RDTR) for the 2m temperature ~~with values near 1, where values greater 1 indicate an overestimation of the diurnal temperature range~~, and the rank correlation coefficient of mean hourly values for precipitation and 10m wind speed. **In case of RDTR a value equal to 1 indicates perfect match, whereas greater than 1 indicates an overestimation of the diurnal temperature range.**”
- Line 293: “CEU-3 GRID”.
- Answer: Will be corrected.
- Lines 297-298: Repetition, could be removed.
- Answer: We will delete the sentence on the indices.
- Lines 311-312: The spread spanned by the station observations is actually not visible in EF. In general, I’d somehow question the usefulness of the station observations in this kind of figure. They surely represent a different spatial scale than the area-averaged model and HYRAS data. Depending on the station network design, the simulated area-average could be outside the station range even in case of a perfect model performance. I certainly also see the value of this additional data source, but the fact that a different spatial scale is represented here should be mentioned at least.
- Answer: We agree that this topic needs a bit more discussion and we will add that information to the manuscript. In our view the station data spread represents the heterogeneity of the area in question. For EF with its flat terrain this is hardly visible and expected. We agree that there could be issues due to an unbalanced station network design. As mentioned above, we will discuss this in more detail by pointing out the potential biases when using this data source.
- Lines 327-328: Where is this shown?
- Answer: This is shown in Figure 4 as mentioned in Line 323 of the same paragraph.

- *Lines 341-343: REMO's unrealistic PDF peak around the freezing point has been known for long, and I actually thought that it has been ameliorated already. Could you provide some brief additional information on the reason for this peak?*
- *Answer: The general issue is freezing and thawing water in the soil. Here we use an approach that considers a temperature range depending on the heterogeneity of the grid box (height and soil type). With increasing resolution, there are more dominant soil type (increasing fraction of one soil type) and the variance of height decreases, because the input data sets we used here did not change compared to the coarser resolved model, where the issue is almost gone. One of the learnings here is that we will use input data at higher resolution and allow for a broader temperature range where freezing/thawing can happen.*
- *Line 348: "offers" sounds a little off here, maybe use "shows" or something similar.*
- *Answer: We agree. We will replace it by "shows".*
- *Lines 352-353: Especially REMO underestimates the DTR. Could this be connected to the unrealistic Tmax/Tmin PDF peaks mentioned before?*
- *Answer: Partly. It seems mostly related to the issue with Tmin as Tmax is represented quite well in REMO.*
- *Figure 5 (and further figures of this kind): It's actually not quite clear to me, which data are shown here as PDF. Is it the regional mean values at daily scale? Or the daily values for all grid cells in the respective region? In the first case, have the station values also been averaged for each day (multi-station averages)? Please clarify. Also, the figure caption does not mention the station dataset.*
- *Answer: It is daily values of all grid cells in the respective region. We will refine this in the text. In addition, we will add the station data to the caption.*
- *Figure 9: The caption says that data for the period 2005-2014 are shown here. However, the TRY reference only extends until 2012. Please check.*
- *Answer: This is correct. We will change that accordingly.*
- *Line 361: I'd rather speak of "spatial pattern" instead of "structure".*
- *Answer: Yes, we agree. Will be changed.*
- *Line 383: I suggest avoiding the term "significantly" if there is no underlying test for statistical significance.*
- *Answer: We will delete "significantly".*
- *Lines 395-396: "it can be observed" à I would avoid this expression in cases where you're referring to simulation data (as here).*
- *Answer: We will change the sentence to: "Moreover, it can be **seen** that the **observed** morning convection minimum, ~~seen in the observational data~~, is often captured."*
- *Figure 14: I suggest relabeling the y-axes (avoid the cryptic e-notation).*
- *Answer: The plots have been changed accordingly (see below).*



- *Lines 407-409: I guess the 8 to 10 minutes shift is caused by the scanning procedure of the sensor. Is there any reference available for this temporal shift of the SARA3 data?*
- *Answer:* The observed temporal offset of approximately 8–10 min is related to the scanning geometry of the geostationary Meteosat SEVIRI instrument. The SARA3 dataset documentation explicitly provides the auxiliary variable `time_difference_SEVIRI`, which accounts for the difference between the nominal product timestamp and the actual image acquisition time for each grid cell (Pfeifroth et al., 2023a). We added the link to the *SARA3 Auxiliary Data User Guide* to the manuscript:
https://public.cmsaf.dwd.de/data/perm/auxilliary_data/SARA3_Aux_Data_User_Guide_V01.pdf
- *Line 424: “subsidence if wind” à I suggest avoiding the term “subsidence” here, since this is probably not what you mean. Replace, for instance, by “the nighttime decrease of wind speed in summer”.*
- *Answer:* Yes, you are right. We will do the suggested replacement.
- *Line 432: “pre-Alpine region PAR”.*
- *Answer:* Will be changed.
- *Line 433 : “Here, all simulations underestimate the wind speed substantially”.*
- *Answer:* Will be changed as suggested.
- *Figure 22: The box plots obviously show interannual variability, i.e. the spread of the 10 annual values. This is not ultimately clear from the figure caption. In general: Why is interannual variability only considered for the evaluation of climate indices, but not in the other cases (e.g., for mean temperature). Biases in the interannual variability of*

mean or max/min temperature could well be the reason for a biased interannual variability of some of the indices.

- Answer: Yes, it is the interannual variability. We suggest adding “Distribution **(interannual variability)** of...” to the caption to make it more explicit. We also agree with the statement that the biases can be connected and suggest making the connection to the earlier results more explicit (see also comment below).
- *Lines 469-472: The mentioned biases of the temperature threshold indices are actually well connected to the biases in tails of the PDFs, as shown before. This could be mentioned here additionally.*
- Answer: Yes, we will make the connection more explicit by adding two more sentences to the paragraph: “Overall, the evaluation of the climate indices indicates a systematic underestimation of heat stress and an apparent overestimation of frost days within the ensemble. **Both results confirm the findings on daily minimum (Tmin) and maximum temperatures (Tmax) discussed above.** The results for precipitation show on average a good representation of R95ptot and an overestimation of R20mm. **This result fits well to the above discussed overestimations in the frequency distribution of daily precipitation amounts.** The ensemble captures the maxima of the number of occurrences for the indices frost days, R20mm, R95ptot in single years. The observed years with maximum heat stress are not necessarily reproduced by the ensemble. We conclude that regional patterns in the different regions assessed are generally reproduced, but also document the need for bias correction as post-processing (Pinto et al., 2026).”
- *Lines 484-497: The qualitative comparison to the results of Kotlarski et al. (2014) are certainly valuable. Not, however, that their underlying ensemble does not only have a coarser spatial grid (12 km) but is also driven by a different re-analysis (ERA Interim). This difference should at least be mentioned. Even better would be some comparison to the performance of the intermediate, ERA5-driven 12 km ensemble (see comment above).*
- Answer: As mentioned above for users of the data, e.g. impact modelers, the reason for the improvement is often not of high priority. We will tackle this issue in a different study and consider it here beyond the scope of the paper.