nature portfolio

Peer Review File

Three decades of simulating global temperature patterns with coupled global climate models

Corresponding Author: Dr Lukas Brunner

This file contains all editorial decision letters in order by version, followed by all author rebuttals in order by version.

Version 0:

Decision Letter:

** Please ensure you delete the link to your author home page in this e-mail if you wish to forward it to your coauthors **

Dear Dr Brunner,

Your manuscript titled "Three decades of simulating global temperatures with coupled global climate models" has now been seen by 2 reviewers, and we include their comments at the end of this message. They find your work of interest, but some important points are raised. We are interested in the possibility of publishing your study in Communications Earth & Environment, but would like to consider your responses to these concerns and assess a revised manuscript before we make a final decision on publication.

We therefore invite you to revise and resubmit your manuscript, along with a point-by-point response that takes into account the points raised. Please highlight all changes in the manuscript text file.

Please submit your point-by-point responses as a separate file, distinct from your cover letter where you can add responses to the Editors' comments that you do not want to be made available to the reviewers. Word files are preferred. We recommend that any figures, tables or graphs that are included in the response to reviewers are also included in the main article or Supplementary Information.

We are committed to providing a fair and constructive peer-review process. Please don't hesitate to contact us if you wish to discuss the revision in more detail.

Please use the following link to submit your revised manuscript, point-by-point response to the referees' comments (which should be in a separate document to any cover letter), a tracked-changes version of the manuscript (as a PDF file) and the completed checklist:
Link Redacted
** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

We hope to receive your revised paper within six weeks; please let us know if you aren't able to submit it within this time so that we can discuss how best to proceed. If we don't hear from you, and the revision process takes significantly longer, we may close your file. In this event, we will still be happy to reconsider your paper at a later date, as long as nothing similar has been accepted for publication at Communications Earth & Environment or published elsewhere in the meantime.

Please do not hesitate to contact us if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Best regards,

ChenRui Diao, PhD

Associate Editor,
Communications Earth & Environment
Consulting Editor,

Communications Sustainability

EDITORIAL POLICIES AND FORMATTING

We ask that you ensure your manuscript complies with our editorial policies. Please ensure that the following formatting requirements are met, and any checklist relevant to your research is completed and uploaded as a Related Manuscript file type with the revised article.

For Manuscripts that fall into the following fields:
• Behavioural and social science
• Ecological, evolutionary & environmental sciences
• Life sciences
An updated and completed version of our Reporting Summary must be uploaded with the revised manuscript
You can download the form here:
https://www.nature.com/documents/nr-reporting-summary.pdf

Furthermore, please align your manuscript with our format requirements, which are summarized on the following checklist: Communications Earth & Environment formatting checklist

and also in our style and formatting guide Communications Earth & Environment formatting guide .

*** DATA: Communications Earth & Environment endorses the principles of the Enabling FAIR data project (http://www.copdess.org/enabling-fair-data-project/). We ask authors to make the data that support their conclusions available in permanent, publically accessible data repositories. (Please contact the editor if you are unable to make your data available).

All Communications Earth & Environment manuscripts must include a section titled "Data Availability" at the end of the Methods section or main text (if no Methods). More information on this policy, is available at http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf.

In particular, the Data availability statement should include:
- Unique identifiers (such as DOIs and hyperlinks for datasets in public repositories)
- Accession codes where appropriate
- If applicable, a statement regarding data available with restrictions
- If a dataset has a Digital Object Identifier (DOI) as its unique identifier, we strongly encourage including this in the Reference list and citing the dataset in the Data Availability Statement.

DATA SOURCES: All new data associated with the paper should be placed in a persistent repository where they can be freely and enduringly accessed. We recommend submitting the data to discipline-specific, community-recognized repositories, where possible and a list of recommended repositories is provided at http://www.nature.com/sdata/policies/repositories.

If a community resource is unavailable, data can be submitted to generalist repositories such as figshare or Dryad Digital Repository. Please provide a unique identifier for the data (for example a DOI or a permanent URL) in the data availability statement, if possible. If the repository does not provide identifiers, we encourage authors to supply the search terms that will return the data. For data that have been obtained from publically available sources, please provide a URL and the specific data product name in the data availability statement. Data with a DOI should be further cited in the methods reference section.

Please refer to our data policies at http://www.nature.com/authors/policies/availability.html.

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

This study makes a valuable contribution to understanding the evolution of global climate model performance over three decades—particularly regarding the potential of km-scale models and the uncertainty associated with reference datasets. The research evaluates 176 coupled global climate models spanning 5 generations (from CMIP2 to km-scale models) over a 30-year period, paired with 10 observation-based reference datasets. This largescale, longterm analysis addresses a gap in existing literature, where studies often focus on single model generations or a limited number of reference datasets. Notably, the study identifies the potential of km-scale models, evidenced by IFS-FESOM outperforming top CMIP6 models, while also highlighting their current limitations, providing critical guidance for future model development. However, several ambiguities remain that the authors should clarify in the revised manuscript. Detailed comments are as follows:

1. A critical oversight is the study's failure to address the "gray zone" challenge of convection parameterization in km-scale models. At resolutions of approximately 1–10 km, deep convection processes are neither fully resolvable nor entirely subgrid-scale. This ambiguity results in inconsistent handling of convection across different km-scale configurations, yet the study does not establish a link between this inconsistency and observed differences in model performance.

2. While the study groups 11 km-scale model configurations together, it lacks in-depth exploration of why specific models outperform others. The connection between model structure (e.g., parameterization schemes, grid design) and performance gaps remains insufficiently elaborated.

3. The exclusive focus on spatial patterns of 2-meter surface air temperature overlooks other critical climate variables (e.g., precipitation) and temporal dynamics (e.g., interannual variability). This narrow scope limits the comprehensiveness of the model evaluation, as a robust assessment should ideally encompass multiple dimensions of climate system behavior.

4. Although the study acknowledges the importance of model tuning, it provides limited details on how specific tuning strategies (e.g., adjustment of top-of-atmosphere radiation balance) interact with "gray zone" convection schemes to enhance performance.

Reviewer #2 (Remarks to the Author):

Review of "Three decades of simulating global temperatures with coupled global climate models" by Brunner et al.

This manuscript presents a remarkably comprehensive and technically compelling assessment of the evolution of coupled global climate models over the past three decades, from CMIP2 to CMIP6 and the first kilometre-scale prototypes, by focusing on the 2-m air temperature (T2m). The analysis framework is rigorous, the dataset impressively broad (176 models and 10 reference datasets), and the presentation carefully executed. The authors convincingly document steady improvements in the simulated T_2m anomaly pattern, the importance of tuning relative to model resolution, and the growing role of reference dataset choice in determining apparent model skill.

I greatly respect the effort and depth of analysis demonstrated in this study. It will serve as a valuable benchmark for the modelling community and provide useful context for CMIP7 and emerging digital-twin initiatives. However, from a scientific perspective, while quantitatively thorough, the results are not especially surprising. The key findings align with what most climate modelers would anticipate: gradual improvement over time, convergence of best practices, limited benefit from higher resolution alone, and an increasing influence of observational uncertainty. Although I am not personally very excited by the novelty, I view this as a solid, worthwhile, and important contribution that merits publication after minor revision.

My major comments are following:

Physical interpretation of improved T2m pattern
The paper's central metric is the improvement in the spatial pattern of 20-year T2m anomaly. While this provides a convenient long-term climatological mean comparison, the physical meaning of this improvement is not clearly articulated. The authors should clarify what processes are responsible for a better temperature pattern—e.g., more realistic land–sea contrast, equator-to-pole gradients, topographic effects, or ocean circulation features.

Without this context, the improvement appears primarily statistical rather than physical. A brief discussion or illustrative figure showing where and why the models have improved, for example, over the North Atlantic, high-latitude continents, or mountainous regions, would make the findings more interpretable and scientifically informative.

Relationship between IFS models and ECMWF operational forecast and reanalyses
The authors highlight that IFS-FESOM outperforms CMIP6 models, demonstrating the potential of kilometre-scale global climate modelling. Based on Figure 1, it appears that entire IFS modeling group performs very well in general. The authors highlight that IFS-FESOM outperforms CMIP6 models, demonstrating the potential of kilometre-scale global climate modelling.

I would like clarification on how closely the IFS-FESOM configuration resembles the operational ECMWF numerical weather prediction system. To what extent does this model benefit from decades of operational tuning and verification? If the IFS-FESOM system inherits much of the ECMWF model physics and tuning, its superior performance relative to CMIP models and ECMWF reanalyses such as ERA5 is perhaps not surprising. Explicitly discussing this connection would help readers interpret the significance of the IFS results and clarify whether they primarily reflect advances in resolution or the advantages of a well-tested, operationally maintained system.

Minor comment:
Figures 1–3 are rich but crowded; larger fonts and succinct caption summaries would help.

** Visit Nature Portfolio's author and referees' website at www.nature.com/authors for information about policies, services and author benefits**

Communications Earth & Environment is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the Manuscript Tracking System by clicking on 'Modify my Springer Nature account' and following the instructions in the link below. Please also inform all co-authors that they can add their ORCIDs to their accounts and that they must do so prior to acceptance.
https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

For more information please visit http://www.springernature.com/orcid

If you experience problems in linking your ORCID, please contact the Platform Support Helpdesk.

Version 1:

Decision Letter:

** Please ensure you delete the link to your author home page in this e-mail if you wish to forward it to your coauthors **

Dear Dr Brunner,

Your manuscript titled "Three decades of simulating global temperature patterns with coupled global climate models" has now been seen by our reviewers, whose comments appear below. In light of their advice we are delighted to say that we are happy, in principle, to publish a suitably revised version in Communications Earth & Environment.

We therefore invite you to revise your paper one last time to address the remaining concerns of our reviewers. At the same time we ask that you edit your manuscript to comply with our format requirements and to maximise the accessibility and therefore the impact of your work.

EDITORIAL REQUESTS:

Please review our specific editorial comments and requests regarding your manuscript in the attached "Editorial Requests Table".

*****Please take care to match our formatting and policy requirements. We will check revised manuscript and return manuscripts that do not comply. Such requests will lead to delays. *****

Please outline your response to each request in the right hand column. Please upload the completed table with your manuscript files as a Related Manuscript file.

If you have any questions or concerns about any of our requests, please do not hesitate to contact me.

SUBMISSION INFORMATION:

In order to accept your paper, we require the files listed at the end of the Editorial Requests Table; the list of required files is also available at https://www.nature.com/documents/commsj-file-checklist.pdf .

OPEN ACCESS:

Communications Earth & Environment is a fully open access journal. Articles are made freely accessible on publication. For further information about article processing charges, open access funding, and advice and support from Nature Portfolio, please visit https://www.nature.com/commsenv/open-access

At acceptance, you will be provided with instructions for completing the open access licence agreement on behalf of all authors. This grants us the necessary permissions to publish your paper. Additionally, you will be asked to declare that all

required third party permissions have been obtained, and to provide billing information in order to pay the article-processing charge (APC).

Please use the following link to submit the above items:
Link Redacted
** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

We hope to hear from you within two weeks; please let us know if you need more time.

Best regards,

ChenRui Diao, PhD

Associate Editor,
Communications Earth & Environment
Consulting Editor,
Communications Sustainability

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

Follow-up Review Comments on "Three decades of simulating global temperatures patterns with coupled global climate models"

The authors have made substantial progress in addressing the review comments. The revised manuscript is scientifically rigorous, well-structured, and impactful. The remaining recommendations are minor and focused on strengthening the physical interpretation and transparency of the analysis. Once these revisions are implemented, the manuscript will be fully suitable for publication. That said, a few minor limitations remain regarding the depth of physical mechanism analysis and the discussion of model structural differences. The manuscript is now nearly publication-ready, and the following suggestions aim to refine rather than substantially revise the work. Below are detailed follow-up comments.

1. The manuscript notes that ICON produces higher localized precipitation while IFS yields better temperature simulations, but it does not explicitly connect these differences to their respective convection treatments or grid configurations.

2. The authors discuss challenges related to convection parameterization but do not explicitly link them to the persistent temperature biases identified. It remains unclear whether the “gray zone” convection treatment in km-scale models exacerbates or alleviates sea ice–related temperature biases. In addition, how differences in convection parameterization influence temperatures in the Arctic marginal ice zone deserves further clarification.

3. My previous comments requested additional details on how tuning strategies interact with convection schemes. Although the authors describe the tuning actions applied, they do not quantify the intensity of tuning for km-scale models, which makes it difficult to disentangle the effects of tuning from those of resolution.

Reviewer #2 (Remarks to the Author):

The authors have thoroughly addressed all my comments and have included additional comprehensive analyses. I am happy to recommend this manuscript for acceptance and publication.

Reviewer #1 (Remarks to the Author):

This study makes a valuable contribution to understanding the evolution of global climate model performance over three decades—particularly regarding the potential of km-scale models and the uncertainty associated with reference datasets. The research evaluates 176 coupled global climate models spanning 5 generations (from CMIP2 to km-scale models) over a 30-year period, paired with 10 observation-based reference datasets. This largescale, longterm analysis addresses a gap in existing literature, where studies often focus on single model generations or a limited number of reference datasets. Notably, the study identifies the potential of km-scale models, evidenced by IFS-FESOM outperforming top CMIP6 models, while also highlighting their current limitations, providing critical guidance for future model development. However, several ambiguities remain that the authors should clarify in the revised manuscript. Detailed comments are as follows:

Thank you to the reviewer for reviewing our manuscript and for the constructive feedback. We have addressed the detailed comments in bold below.

1. A critical oversight is the study's failure to address the "gray zone" challenge of convection parameterization in km-scale models. At resolutions of approximately 1–10 km, deep convection processes are neither fully resolvable nor entirely subgrid-scale. This ambiguity results in inconsistent handling of convection across different km-scale configurations, yet the study does not establish a link between this inconsistency and observed differences in model performance.

Thanks to the reviewer for raising this issue, which is indeed a crucial development in current km-scale models. In the revised manuscript, we have added a paragraph discussing the treatment of convection parameterization and its implications to the Summary and Discussion section.

For your convenience, the relevant part of the new paragraph is also copied below. (The paragraph starts at line 292 in the revised manuscript, the relevant part in line 298):

"*[..] At the same time, the representation of (extreme) precipitation differs quite considerably between the models, with ICON generally showing higher and temporally and spatially more localized precipitation than IFS (Brunner, Poschlod, et al. 2025; Spät et al. 2024; Takasuka, Becker, and Bao n.d.; Wille et al. 2025). This can be attributed to the different treatments of convection in the two models as their grid spacing of about 10 km puts them in somewhat of a transition zone which is slightly too coarse to accurately permit explicit convection (often set to about 2 km to 4 km; e.g., Prein et al. 2021) but too fine to allow a robust empirical parameterization (about 25 km to 50 km; Vergara-Temprado et al. 2020). While our results for mean temperature show that the parameterized runs from IFS tend to perform better than the ICON runs, this result cannot be attributed to the treatment of convection alone and might be due to a whole range of structural model differences. In fact, recent work has shown that a realistic precipitation climatology can be achieved with explicit convection (Daisuke et al. 2024) and that land-atmosphere coupling is better represented, affecting (extreme) temperatures (Lee et al. 2024). Ongoing efforts to run climatological time-periods at 5 km grid spacing globally (Doblas-Reyes et al. 2025), contributing to push km-scale models towards the explicit convection domain, hence, further contribute to materializing their potential.*"

2. While the study groups 11 km-scale model configurations together, it lacks in-depth exploration of why specific models outperform others. The connection between model structure (e.g., parameterization schemes, grid design) and performance gaps remains insufficiently elaborated.

We acknowledge that detailed model descriptions are important to pinpoint the underlying reasons for performance differences, while they are also necessarily limited in an overview study like the one presented. We also note that several of the km-scale model versions are still prototypes and, for example, expertise in how to best tune them is still being developed, while modelling centers might have many years of experience in tuning CMIP-type models. Model performance, hence, arises from a complex interplay between many moving parts, including but not limited to resolution, parameterizations, and tuning (see also figure 3 and related discussions).

As such, our main aim in this work is to showcase the *potential* of km-scale models and to put them into a long-term context (at least for the case of surface air temperature), not to detail their development process and track related performance (which would anyway be limited by our approach of focusing only on temperature). Once these km-scale models arrive at a more consolidated state, it is definitely worth investigating their structural differences to established CMIP models and to each other, and the implications for model output performance.

We provide an overview of the most important km-scale model features in the supplement of the original submission. In the revised manuscript, we have slightly extended this description and moved it to the methods section to increase its visibility (line 349 onward). For a more detailed description of the overall model layout, we refer to the literature existing so far (Hohenegger et al. 2023; Rackow et al. 2025), in particular to the recent work by Segura et al. (2025), where the evolution of ICON and IFS versions during their development is discussed.

3. The exclusive focus on spatial patterns of 2-meter surface air temperature overlooks other critical climate variables (e.g., precipitation) and temporal dynamics (e.g., interannual variability). This narrow scope limits the comprehensiveness of the model evaluation, as a robust assessment should ideally encompass multiple dimensions of climate system behavior.

We agree with the reviewer that the focus on surface air temperature is a clear limitation of our work. As a result, our work is not intended to provide a comprehensive model evaluation, but rather focuses on a single variable. Yet, at the same time, this focus enables us to provide the unique long-term analysis we present in the manuscript, as detailed in the introduction:

Line 43: "*While focusing on mean temperature pattern does not provide a comprehensive model evaluation (which is covered in other work, e.g., Bock et al. 2020) but rather focuses on a narrow aspect of model performance, this has two main advantages: First, it allows for a long-term view on model performance, including early coupled models (which are limited in available variables and provide only simulations of a stable climate) all the way to the latest km-scale models (which are limited in available years). Second, our temperature metric provides a robust starting*

***point, as it can be compared to a range of high-quality, observation-based products. We use this availability of multiple reference datasets to quantify the effect of using different references and to compare model performance with cross-evaluated references performance.*"**

We have also added a sentence to Summary and Discussion section highlighting the focus of our study again:
Line 280: "*Crucially, we note that while focusing solely on surface temperature has enabled the long-term multi-model, multi-reference perspective we provide, it necessarily also omits many climate system features important for a more comprehensive model evaluation.*"

4. Although the study acknowledges the importance of model tuning, it provides limited details on how specific tuning strategies (e.g., adjustment of top-of-atmosphere radiation balance) interact with "gray zone" convection schemes to enhance performance.
While the reviewer clearly raises an important point, relevant to understanding the effect of tuning at different resolutions, we note that our study has a different focus and aims to provide a high-level overview of the model representation of the mean temperature pattern across model generations.

With regard to km-scale models in general (representing the class of models most notably approaching the 'gray zone'), tuning strategies have indeed been adapted to how convection is represented as discussed, for example, in Hohenegger et al. (2020). More specifically, the tuning efforts for ICON and IFS within nextGEMS are described in more detail in Segura et al. (2025; see their section 4).

To address this comment, we have added some additional discussion on the differences between the two km-scale models used (ICON and IFS), which differ in their treatment of convection and can, hence, shed some light on its effects (starting line 292; see also answer to comment 1).
We have also moved the description of the different km-scale model versions and their tuning strategies from the supplement to the methods section to increase visibility (starting line 349; see also answer to comment 2).

Reviewer #2 (Remarks to the Author):

Review of “Three decades of simulating global temperatures with coupled global climate models” by Brunner et al.

This manuscript presents a remarkably comprehensive and technically compelling assessment of the evolution of coupled global climate models over the past three decades, from CMIP2 to CMIP6 and the first kilometre-scale prototypes, by focusing on the 2-m air temperature (T2m). The analysis framework is rigorous, the dataset impressively broad (176 models and 10 reference datasets), and the presentation carefully executed. The authors convincingly document steady improvements in the simulated T_2m anomaly pattern, the importance of tuning relative to model resolution, and the growing role of reference dataset choice in determining apparent model skill.

I greatly respect the effort and depth of analysis demonstrated in this study. It will serve as a valuable benchmark for the modelling community and provide useful context for CMIP7 and emerging digital-twin initiatives. However, from a scientific perspective, while quantitatively thorough, the results are not especially surprising. The key findings align with what most climate modelers would anticipate: gradual improvement over time, convergence of best practices, limited benefit from higher resolution alone, and an increasing influence of observational uncertainty. Although I am not personally very excited by the novelty, I view this as a solid, worthwhile, and important contribution that merits publication after minor revision.
Thank you to the reviewer for reviewing our manuscript and for the constructive feedback. We have addressed the detailed comments in bold below.

My major comments are following:

Physical interpretation of improved T2m pattern
The paper’s central metric is the improvement in the spatial pattern of 20-year T2m anomaly. While this provides a convenient long-term climatological mean comparison, the physical meaning of this improvement is not clearly articulated. The authors should clarify what processes are responsible for a better temperature pattern—e.g., more realistic land–sea contrast, equator-to-pole gradients, topographic effects, or ocean circulation features.

Without this context, the improvement appears primarily statistical rather than physical. A brief discussion or illustrative figure showing where and why the models have improved, for example, over the North Atlantic, high-latitude continents, or mountainous regions, would make the findings more interpretable and scientifically informative.
Thank you to the reviewer for this suggestion. We have included an additional figure (figure S2), revealing the regions of main model improvements, and added additional discussion.

For your convenience, we have included the figure below. It shows the development of bias across CMIP and km-scale models, as detailed in the caption (please note that a database of distance maps for each individual model can also be found in the online supplement at https://cloud.uni-hamburg.de/s/g27Mp3AN8q2C4Nc).

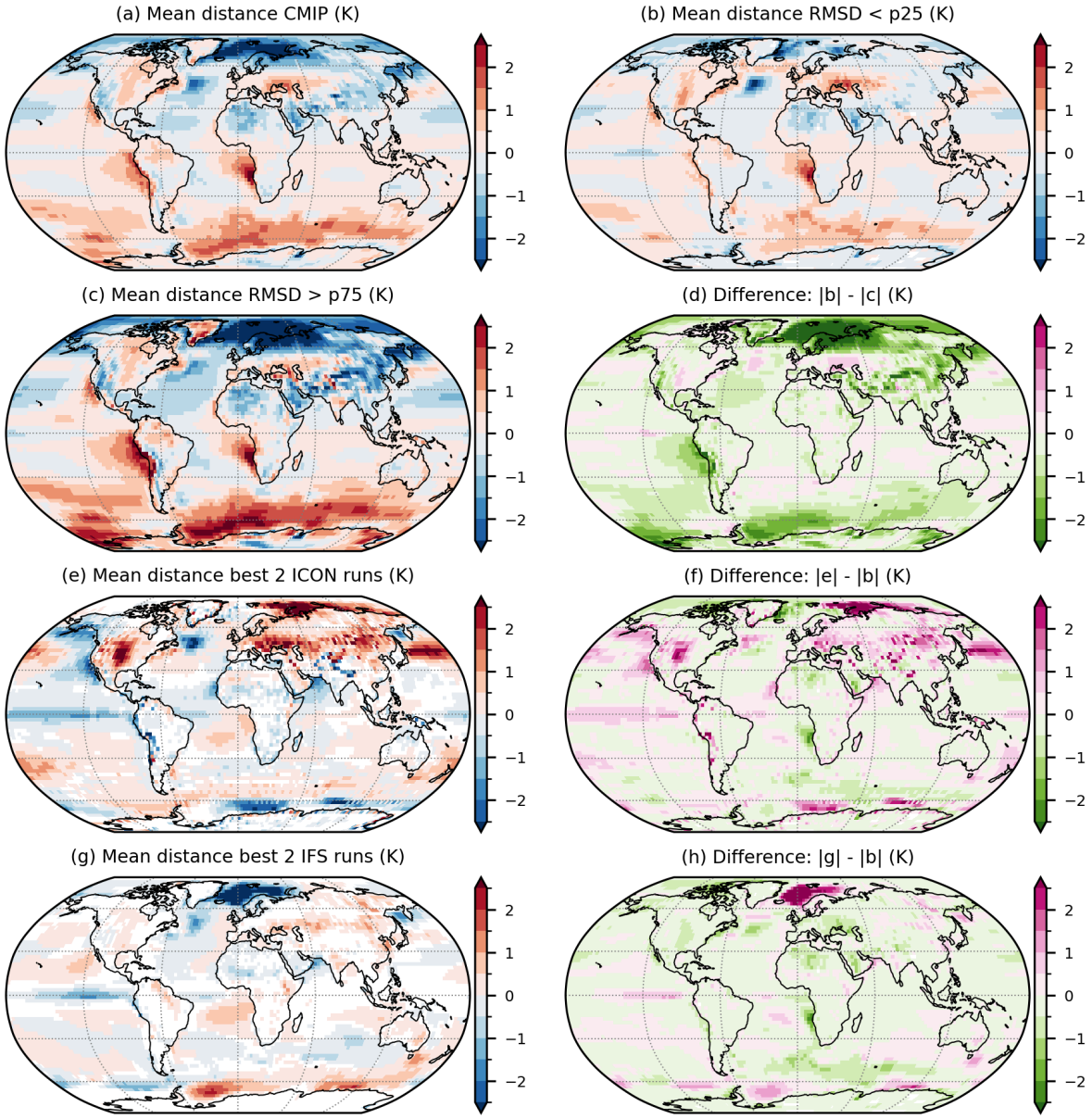

Figure S2: Multi-model mean bias for (a) all 165 CMIP models, (b) the 25 % models with the lowest RMSD (41 models), (c) the 25 % models with the highest RMSD, and (d) the
difference of the absolute values of (b) and (c). Mean of the two best km-scale runs for the (e) ICON and (g) IFS model, as well as the absolute value of differences to the case shown in (c).

The related discussion starts in line 125 in the revised manuscript and is copied in below.
"*More generally, the northern North Atlantic is a region with quite persistent multi-model mean bias, as the spatially resolved mean distance across all 165 CMIP models in figure S2a and the individual model bias maps in the online supplement show. Similarly, many models have a warm bias in the Southern Ocean that also persists through generations. The northern hemisphere cold bias mainly stems from*

the winter season and has been connected to the representation of sea ice in the models (e.g., Davy 2020), while the southern hemisphere warm bias has been connected to remote effects (e.g., Luo et al. 2023). Yet another set of regions with particularly persistent temperature biases can be found in the low cloud regions (e.g., Chen et al. 2022) at the eastern edges of the ocean basins, with model deficiencies documented already for the first CMIP (Meehl, Boer, et al. 1997) still persisting in CMIP6 (Bock et al. 2020) but starting to be resolved in the km-scale models (figures S2e,g). While pointing at systematic problems in the models, such high-bias regions naturally also have the highest potential for improvements if the mechanisms underlying the biases are understood and can be addressed in model development. This potential could (at least partly) be realized throughout CMIP and in the km-scale models (figure S2b-h). Large improvements can be seen between the top and bottom performing CMIP models, in particular at high latitudes. The km-scale models in general and IFS in particular are able to improve further, beyond the top quarter of CMIP models across many regions (figure S2f,h).”

Relationship between IFS models and ECMWF operational forecast and reanalyses
The authors highlight that IFS-FESOM outperforms CMIP6 models, demonstrating the potential of kilometre-scale global climate modelling. Based on Figure 1, it appears that entire IFS modeling group performs very well in general. The authors highlight that IFS-FESOM outperforms CMIP6 models, demonstrating the potential of kilometre-scale global climate modelling.

I would like clarification on how closely the IFS-FESOM configuration resembles the operational ECMWF numerical weather prediction system. To what extent does this model benefit from decades of operational tuning and verification? If the IFS-FESOM system inherits much of the ECMWF model physics and tuning, its superior performance relative to CMIP models and ECMWF reanalyses such as ERA5 is perhaps not surprising. Explicitly discussing this connection would help readers interpret the significance of the IFS results and clarify whether they primarily reflect advances in resolution or the advantages of a well-tested, operationally maintained system.
The reviewer is correct, the “climate versions” of the km-scale IFS are inherited from the operational numerical weather prediction (NWP) model (see Rackow et al. 2025 for the exact versions). The NWP-IFS has, indeed, been tuned for performance in numerical weather forecasting for several decades. Yet, while climate-IFS runs over long timescales (as for this work) may benefit from this ancestry, other modern climate models also often build on several decades of model development.

In general, the required changes for running IFS over climate timescales have been considerable, as it is not a priori clear how skill in numerical weather forecasting translates into skill in simulating climate. The operational NWP-IFS is not tuned to be long-term stable, conservation properties (mass/energy) are not of crucial concern over medium-range weather prediction timescales, and the coupling with another ocean model (FESOM) meant that optimized settings, e.g., in terms of air-ocean flux exchanges developed for the weather context, may not work for the IFS-FESOM model out of the box. After a change of the ocean component, typically, the atmospheric model components in climate models need to be re-tuned at the top of the atmosphere (TOA). For IFS, the tuning of the TOA radiation balance has been

performed in short IFS-NEMO runs over 12 days, in order to bring the short- and longwave components of the TOA balance close to satellite observations. For the long climate runs with IFS-FESOM at 4km resolution, the identical values determined from the short IFS-NEMO forecast experiments at 4km resolution could be used.

Another point is that the move to 4km resolution in IFS-FESOM required substantial changes to the treatment of deep convection. Introducing a reduced cloud base mass flux for the 4km version is a large step away from the 9km NWP-IFS (which uses a full parameterization for deep convection). In summary, we can say that IFS-FESOM, of course, inherits many aspects from the NWP system, but the move to 4km resolution together with another ocean model results in a substantially different model configuration. For more details on the model versions used and the changes, we refer to the reference publication by Rackow et al. (2025).

In the manuscript a brief description of the connection of weather and climate IFS can be found from line 174 onward: "*While the operational IFS model is coupled to the NEMO ocean model, for the purpose of several climate applications the IFS has also been coupled to the FESOM ocean model and been tuned in two ways: (1) top-of-the-atmosphere (TOA) radiation balance has been tuned to match satellite observations and (2) a reduced cloud base mass flux has been implemented that impacts the precipitation distribution, effectively tuning intense precipitation towards observations (see Rackow et al. 2025, for details). The latter was not done for IFS EERIE-p1.*"

While the above addresses dependence from a structural point (i.e., what are overlaps in the source code), we also note that, from an output perspective, all km-scale IFS model versions are closest to ERA5 out of the 10 reference datasets used (see figure 2c for an example and table S5 for a full list). Based on these results, comparing the climate-IFS versions to NWP-IFS (i.e., ERA5) does not provide a robust assessment of model performance and leads to an advantage over other models in a multi-model comparison. This possibility for dependencies of models and references is one reason for our recommendation to use more than one reference dataset.

See, e.g., line 87 of the revised manuscript: "*[Using multiple references] avoids the potential of underestimating model error due to dependencies between individual models and references [...].*" or line 324: "*Ultimately, we recommend to include reference uncertainty in model evaluations whenever possible, to document model-reference dependencies arising from the development cycle, and to account for these dependencies model evaluations.*"

Minor comment:
Figures 1–3 are rich but crowded; larger fonts and succinct caption summaries would help.
Thanks to the reviewer for pointing this out. We have increased the fontsize of all figures, shortened the caption texts, and added additional titles to the panels in figure 2.

Reviewer #1 (Remarks to the Author):

Follow-up Review Comments on "Three decades of simulating global temperatures patterns with coupled global climate models"

The authors have made substantial progress in addressing the review comments. The revised manuscript is scientifically rigorous, well-structured, and impactful. The remaining recommendations are minor and focused on strengthening the physical interpretation and transparency of the analysis. Once these revisions are implemented, the manuscript will be fully suitable for publication. That said, a few minor limitations remain regarding the depth of physical mechanism analysis and the discussion of model structural differences. The manuscript is now nearly publication-ready, and the following suggestions aim to refine rather than substantially revise the work. Below are detailed follow-up comments.

Thank you to the reviewer for the positive assessment of our revisions and the final suggestions. We have addressed them below in bold.

1. The manuscript notes that ICON produces higher localized precipitation while IFS yields better temperature simulations, but it does not explicitly connect these differences to their respective convection treatments or grid configurations.

Thank you to the reviewer for pointing this out; this was indeed not clear from the text in the manuscript. We have now updated this to read:

“At the same time, the representation of (extreme) precipitation differs quite considerably between the models, with ICON generally showing higher, and temporally and spatially more localized, precipitation than IFS. The higher localization of precipitation in ICON is a result of its setup without any convection parameterization, while IFS still uses parameterization to different degrees (Brunner, Poschlod, et al. 2025; Spät et al. 2024; Takasuka, Becker, and Bao n.d.; Wille et al. 2025; see also table S2).”

For additional details, we might refer the reviewer to the (recently accepted) pre-print by Takasuka et al., currently already available from the ESS open archive (https://essopenarchive.org/doi/full/10.22541/essoar.174834999.937887210), in particular figure S2a in the supplement. It compares IFS run at 9km, with both convection parameterization on and off. Without parameterization, IFS also becomes more similar to ICON and NICAM, which are also run without parameterization (compare supplement figure S2a and figure 3a in the main manuscript of Takasuka et al.).

2. The authors discuss challenges related to convection parameterization but do not explicitly link them to the persistent temperature biases identified. It remains unclear whether the “gray zone” convection treatment in km-scale models exacerbates or alleviates sea ice–related temperature biases. In addition, how differences in convection parameterization influence temperatures in the Arctic marginal ice zone deserves further clarification.

The reviewer raises an interesting question, yet without dedicated experiments (e.g., running an ensemble of km-scale simulations with different treatments of convection), this aspect is impossible to conclusively link to sea ice biases.

We assume that the comment is referring to the sea bias in several of the IFS runs (as shown, e.g., in figure 1f). We might refer to Rackow et al. (2025, in particular section 3.2.4) for some additional information. As they discuss, the sea ice bias emerges in IFS coupled to both the NEMO and FESOM ocean modules. Rather than to the treatment of convection, the majority of the bias is, most likely, connected to the process of coupling of atmosphere, ocean and ice, which is not unexpected given the rather short spinup times (made necessary by the resource usage of the km-scale models). Other work, based on the Destination Earth "nudged storylines" with IFS-FESOM (the IFS configurations used here are identical to the configuration used in the Destination Earth), has linked the observed sea ice biases to the sea ice thermodynamics itself (John et al., 2026; https://doi.org/10.22541/essoar.173160166.64258929/v2), successfully isolating limitations in model physics from differences due to the different evolution of the large-scale atmosphere in free-running simulations. One focus of model development for the IFS will therefore be to guarantee a more consistent treatment of sea ice thermodynamics between the ocean and atmosphere components.

3. My previous comments requested additional details on how tuning strategies interact with convection schemes. Although the authors describe the tuning actions applied, they do not quantify the intensity of tuning for km-scale models, which makes it difficult to disentangle the effects of tuning from those of resolution.

As detailed in our answers to the first round of reviews, we agree that quantifying the effect of tuning and disentangling it from the effect resolution is a relevant endeavour, but out of scope for this study.

The focus of our study is on a high-level overview of the model representation of mean temperature across more than 150 coupled climate models and across multiple decades. To comprehensively assess the effect of tuning in this setting would require information on the tuning procedure, not only for the km-scale models, but for all CMIP models. A true quantification would even require dedicated experiments, varying the tuning strength, to track the effects. Unfortunately, such an effort is not feasible for us.

Reviewer #2 (Remarks to the Author):

The authors have thoroughly addressed all my comments and have included additional comprehensive analyses. I am happy to recommend this manuscript for acceptance and publication.

Thank you to the reviewer for the positive assessment of our revisions and for taking the time.