communications earth & environment

Article

A Nature Portfolio journal

https://doi.org/10.1038/s43247-026-03497-w

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

Check for updates

Lukas Brunner[1,2] ✉, Rohit Ghosh[3], Leopold Haimberger[2], Cathy Hohenegger[4], Dian Putrasahan[4], Thomas Rackow[5], Reto Knutti[6] & Aiko Voigt[2]

Accurately simulating the climate system has been a challenge and aspiration since the advent of numerical modeling. Here, we use the spatial pattern of 2 m surface temperature to discuss the evolution of model performance from the beginning of the Coupled Model Intercomparison Project (CMIP) in the 1990s to the latest kilometer-scale models today. We find that the kilometer-scale IFS-FESOM model outperforms even the best CMIP6 models, while other kilometer-scale models still have considerable deficits. These results demonstrate the potential of kilometer-scale models to surpass established CMIP models, despite undergoing only limited tuning, while also highlighting the considerable efforts still needed to realize their full potential. We put this performance in the context of 10 observation-based references and 150 coupled global climate models developed over the past three decades to discuss that increasing resolution might be necessary, but is not sufficient for improving model skill.

Today's global coupled climate models build on a long history of development, from the first attempts to couple atmospheric and ocean models in the late 1960s[1], and the ability to simulate a stable climate without flux adjustments in the 1990s[2], to the emergence of Earth system models in the 2000s[3], and the development of the first kilometer-scale (km-scale) models today[4–6]. Over the decades, the performance of models to represent Earth's climate has improved, driven by increases in computing power, growing knowledge of and better observing systems for the climate system, and enhanced understanding of small-scale processes[7–9]. These developments have allowed to run increasingly complex models at higher resolutions and have reduced the reliance on process omission, parameterization, and empirical correction for processes that cannot be fully represented[4,10–12]. Yet, all models are approximations of the real climate system, and their output has to be compared to observation-based references to ensure fitness for a given purpose or reveal inadequacies[13–16].

In fact, the understanding that model evaluation and comparison is crucial has led to the first Coupled Model Intercomparison Project (CMIP) in the 1990s, whose main objective was to "*document systematic simulation errors of global coupled climate models […] by comparing the mean model output to observations to determine how well the coupled models simulate current mean climate.*"[17] Here, we build on this and use the spatial pattern of 20-year mean 2 m surface air temperature to track the evolution of model performance over three decades, from the advent of CMIP to today. We evaluate 176 fully coupled models against a set of 10 observation-based reference datasets. While focusing on the mean temperature pattern does not provide a comprehensive model evaluation [which is covered in other work, e.g., ref. 18] but rather focuses on a narrow aspect of model performance, it 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.

Since its establishment in the 1990s, CMIP has provided the framework for consistently comparing models[17]. However, many model evaluation studies, impact assessments, and high-level reports, including those from the Intergovernmental Panel on Climate Change, focus only on the latest generation of models available at the respective time [e.g., ref. 19]. Studies comparing models across generations, in contrast, often take a comprehensive approach and investigate a large suite of variables, aggregated into

[1]Research Unit Sustainability and Climate Risk, Earth and Society Research Hub (ESRAH), University of Hamburg, Hamburg, Germany. [2]Department of Meteorology and Geophysics, University of Vienna, Vienna, Austria. [3]Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research (AWI), Bremerhaven, Germany. [4]Max Planck Institute for Meteorology, Hamburg, Germany. [5]European Centre for Medium-Range Weather Forecasts (ECMWF), Bonn, Germany. [6]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland. ✉e-mail: lukas.brunner@uni-hamburg.de

high-level metrics such as multi-model means and global means. In addition, they often rely on a single observation-based dataset as a reference [e.g., refs. 7,18,20,21]. While multi-model means can have the advantage to average out different model-dependent biases, such an approach tends to conceal model diversity within model generations, hides spatial performance differences across models, and assumes that a single observational reference is sufficient to serve as a baseline.

We showcase and discuss the implications of these points by drawing on an extensive archive of 176 models, developed over three decades and five model generations: 15 models from the 2nd phase of CMIP (CMIP2; refs. 2,22), 22 from CMIP3[23,24], 44 from CMIP5[25], 84 from CMIP6 (including 17 from the High Resolution MIP—HighResMIP; refs. 26,27), and 11 runs from two next-generation km-scale models[4,5]. Note that global temperature fields were not available from CMIP1[17]. We evaluate the models against 10 observation-based references, among them several historically relevant but nowadays superseded datasets: 20CR[28], Berkeley Earth[29], ERA40[30], ERA-Interim[31], ERA5[32], JRA55[33], JRA3Q[34], MERRA[35], MERRA2[36], and NCAR-NCEP[37].

We use a standard evaluation metric, adapted to allow the option of incorporating multiple references. At each grid cell, the distance of the 20-year mean temperature anomaly relative to the global mean between a model i ($\tau_{\mathrm{model\_i}}$) and one or more references (τ_{ref}) is calculated following:

$$\Delta\tau_{\mathrm{model\_i}} = \begin{cases} \tau_{\mathrm{model\_i}} - \tau_{\mathrm{ref\_min}} & \text{if } \tau_{\mathrm{model\_i}} < \tau_{\mathrm{ref\_min}} \\ \tau_{\mathrm{model\_i}} - \tau_{\mathrm{ref\_max}} & \text{if } \tau_{\mathrm{model\_i}} > \tau_{\mathrm{ref\_max}} \\ 0 & \text{else}, \end{cases} \quad (1)$$

vice-versa. This approach (1) accounts for the (historical) uncertainty in our knowledge of the true state of the climate system by assigning a distance $\Delta\tau_{\mathrm{model\_i}}$ of zero if model i is within the references at a given grid cell, (2) avoids the potential of underestimating model error due to dependencies between individual models and references, and (3) allows for a comparison of model performance against reference performance in a leave-one-out cross-validation that compares one reference against the others.

20-year mean values are calculated on a common 2.5° × 2.5° grid from the models' pre-industrial control simulations (or the closest corresponding run if pre-industrial is not available; a few km-scale models with less than 20 years were also included; see "Methods" for details). From the bias fields (Fig. 1e–g), we calculate the area-weighted root-mean-square distance (RMSD) as well as the fraction of Earth's surface where a model lies within the reference range ($\Delta\tau_{\mathrm{model\_i}} = 0$; see "Methods" for details).

Results

Simulating the global temperature pattern from the beginning of CMIP until km-scale models today

The fraction of Earth's surface for which a given model simulates surface temperature within the bounds of the 10 references has increased continuously from a median of 25% in CMIP2 to 36% in CMIP6 (Fig. 1a). Yet, a clear gap remains between the best-performing CMIP model (GFDL-CM4: within the reference range for about half the globe—49%) and the most distinct cross-evaluated reference (Berkeley Earth: 63%). However, several of the next-generation km-scale model runs set new records for this metric, as they simulate temperature climatologies within the reference range for up to 60% (IFS-FESOM EERIE) of Earth's surface for the first time and come close to the lower end of the cross-evaluated references. When focusing only on the land surface, two of the IFS-FESOM runs even outperform several cross-validated references (Fig. 1c; see Fig. S3 for ocean only).

This finding is yet more pronounced for the land RMSD, where only four out of the 10 references (ERA40, ERA5, JRA55, JRA3Q) perform better than the best IFS-FESOM runs, while ERA-Interim ranks similar and the other references, including MERRA2, a state-of-the-art reanalysis, are outperformed (Fig. 1d). When including the ocean, there is still a gap between the RMSD performance of the best models and the cross-validated references (Fig. 1b), and the best km-scale models do not perform better than the coarser CMIP6-generation models. The differences can be explained by the fact that the RMSD (shown in Fig. 1b, d) also considers the amplitude of the bias while the fractional metric (Fig. 1a, c) does not. While several of the IFS model versions perform very well overall, they are affected by a strong negative temperature bias in the northern North Atlantic (see, e.g., Fig. 1f), affecting their RMSD score. This bias is related to excessive Arctic sea ice formation over the marginal ice regions and is potentially linked to the sea ice coupling scheme. The issue is currently being addressed by implementing a more dynamic coupling approach, and we expect global RMSD to decrease once this issue is solved.

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 Fig. S2a and the individual model bias maps in the online supplement show[38]. 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., ref. 39], while the southern hemisphere warm bias has been connected to remote effects [e.g., ref. 40]. Yet another set of regions with particularly persistent temperature biases can be found in the low cloud regions at the eastern edges of the ocean basins [e.g., ref. 40], with model deficiencies documented already for the first CMIP (Meehl et al.[17]) still persisting in CMIP6 (Bock et al.[18]) but starting to be resolved in the km-scale models (Fig. 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 as shown in Fig. S2b–h. Large improvements in the representation of surface temperature can be seen between the top and bottom performing CMIP models (Fig. S2d), and the km-scale models in general and IFS in particular are able to improve this further, beyond the top quarter of CMIP models (Fig. S2f, h).

To estimate the influence of the remaining internal variability in the 20-year mean temperature field used as a performance metric, we also show the distribution of a 50-member Single Model Initial-condition Large Ensemble (SMILE) from the CMIP6-generation MPI-ESM1-2-LR model[41]. The range of the 50 SMILE members is generally quite small compared to the differences in the respective metrics described above. For the case of the global RMSD, the full range amounts to 0.84 to 0.95 K or about 9% of the CMIP median RMSD. A collection of bias maps for all 176 models, 50 SMILEs, and 10 references is provided in the online supplement. In addition, for models with sufficiently long pre-industrial control runs, internal variability is estimated from additional 20-year slices (Fig. S11).

In a historical context, we find that the ability of CMIP models to simulate surface air temperature patterns has not consistently improved when considering the best-performing model of each generation. Although the multi-model median shows a clear improvement, consistent with findings in the literature [e.g., ref. 18], this appears to rather reflect a reduction in poorly performing models and a convergence toward best practices, rather than a true advancement in best model performance. In the case of the global inside area fraction (Fig. 1a), the best CMIP6 model (GFDL-CM4; 49%) performs similarly to the best CMIP2 model (UKMO; 46%; although the highest values are only achieved in flux-adjusted models for CMIP2). Similarly, for the RMSD over land between CMIP5 and CMIP6 (Fig. 1d). In this context, we stress that the aim of model development, ultimately, has to be to improve not the multi-model mean but the best models in as many metrics as possible.

In summary, we have shown that km-scale models can shift the boundaries of what is achievable. However, while our results demonstrate the potential of km-scale models to outperform coarser models, running models at the km-scale alone is no guarantee for success[42]. In fact, most of the current km-scale model runs do not perform better than traditional CMIP models in simulating 2 m surface air temperature patterns. This is, perhaps, not surprising given that they are still in a prototype state and have

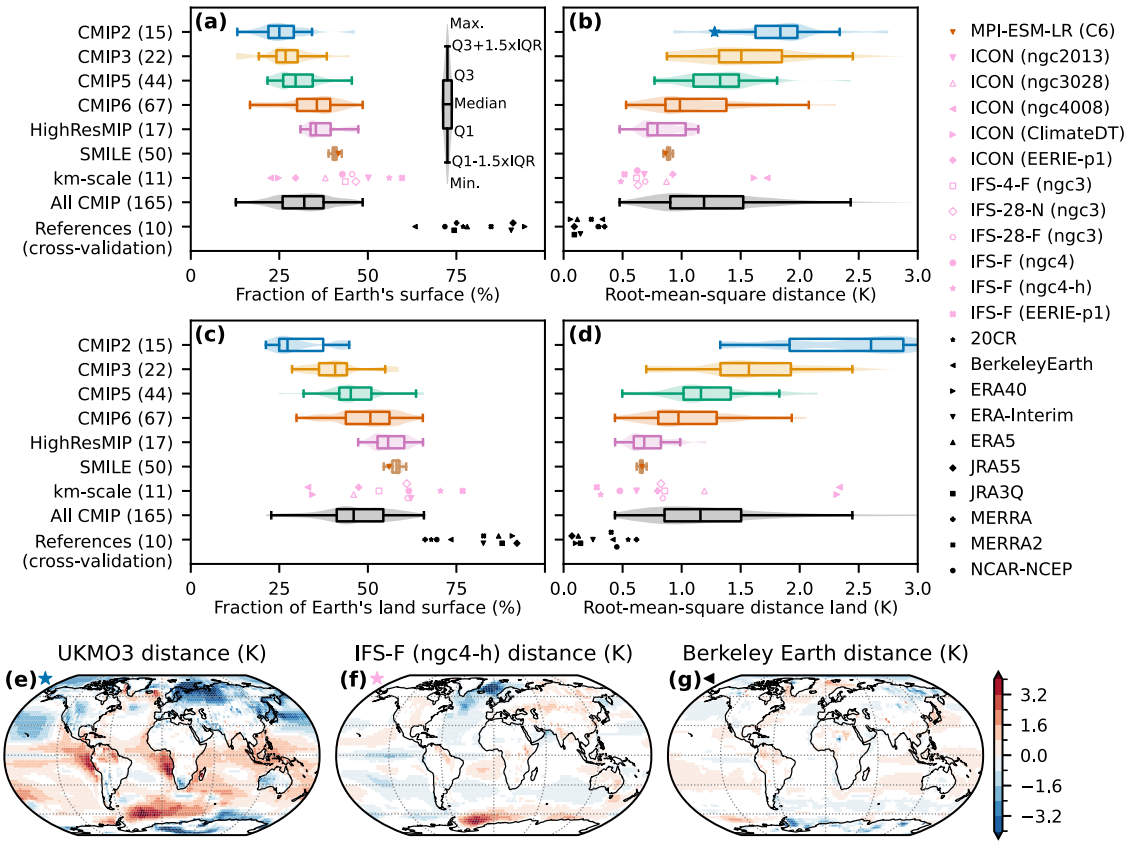

Fig. 1 | Evolution of model performance across generations. a Fraction of Earth's surface area for which the models' 20-year mean surface temperature falls within the range of the 10 references, and **b** area-weighted root-mean-square distance (RMSD) relative to the reference range. **c**, **d** same as (**a**, **b**) but only for the land surface. Unfilled symbols indicate runs with less than 20 years available. **e**–**g** Distance to the reference range, with grid cells where the model falls between the references shown in white. The examples show the models with the lowest RMSD in (**e**) CMIP2 without flux adjustment and (**f**) overall, as well as (**g**) the reference dataset Berkeley Earth, cross-evaluated against the other nine references.

only been tuned very rudimentarily (for climate applications)[43]. To give an example: the base version of IFS has been continuously improved over several decades for the best performance in terms of numerical weather prediction. 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 ref. 5, for details]. The latter was not done for IFS EERIE-p1. For ICON, only the TOA radiation balance has been tuned, except for the EERIE-p1 version of ICON, where no tuning has been done at all. Since several of the km-scale models are prototype versions or still in development, only limited documentation on their technical specifications exists[4,5]. We, hence, provide a summary of their main features in the "Methods" section and discuss the main advances over coarser models in the final section of this manuscript.

Our approach of using the full range of 10 references as the baseline range for model evaluation might be considered to overestimate uncertainty in our baseline. Yet, our primary rationale for including these reference datasets is that each has been considered state-of-the-art at its time and has been used as a baseline for model evaluation in the literature (for example, ERA40 before the introduction of ERA-Interim and the subsequent ERA5). Importantly, all conclusions presented here hold up under several robustness checks, including when restricting the analysis to the three state-of-the-art reanalyses: ERA5, JRA3Q, and MERRA2 (Fig. S4). Even more crucially, we show below that, when using only a single reference, the choice of dataset can lead to model errors that differ by as much as 40% and lead to a systematic effect depending on the model generation. This clearly shows that overall model performance has reached a point where a single observation-based reference can no longer adequately serve as a reliable reference. So far, this argument has only been made qualitatively [e.g., ref. 44]; here, we provide the first quantification.

Quantifying the effect of reference choice on model performance

Comparing model performance evaluated against each of the 10 reference datasets individually reveals that up to 40% of model RMSD can result from the choice of the reference for the latest km-scale models (Fig. 2a, b). This approach corresponds to applying Eq. (1) 10 times (once for each reference) and then calculating the full range of possible RMSD values divided by their mean value. Studies drawing on only a single reference disregard the differences we reveal here, more or less implicitly assuming that model bias dominates in model evaluations [e.g., ref. 21].

This assumption is defensible for early models, with the effect of reference being mostly constrained to less than 20% for CMIP2, CMIP3, and CMIP5 (Fig. 2a). However, for CMIP6-generation models, the choice of reference can already lead to RMSD variations of up to 30%, increasing to 40% for the latest km-scale models. The main driver of this increase is the reduction of absolute RMSD, with the effect of reference choice staying mostly constant in absolute terms.

Yet the reduction in absolute RMSD is not the sole driver and other model properties also play a role: the model with the strongest dependence of temperature RMSD on the choice of reference is IFS-28-N, which is not the best-performing model in terms of overall RMSD (Fig. 2b). Its sensitivity is, hence, a combination of (1) its low mean RMSD, (2) it being quite close to ERA5 (3rd lowest RMSD to ERA5 model overall), and at the same time (3) it

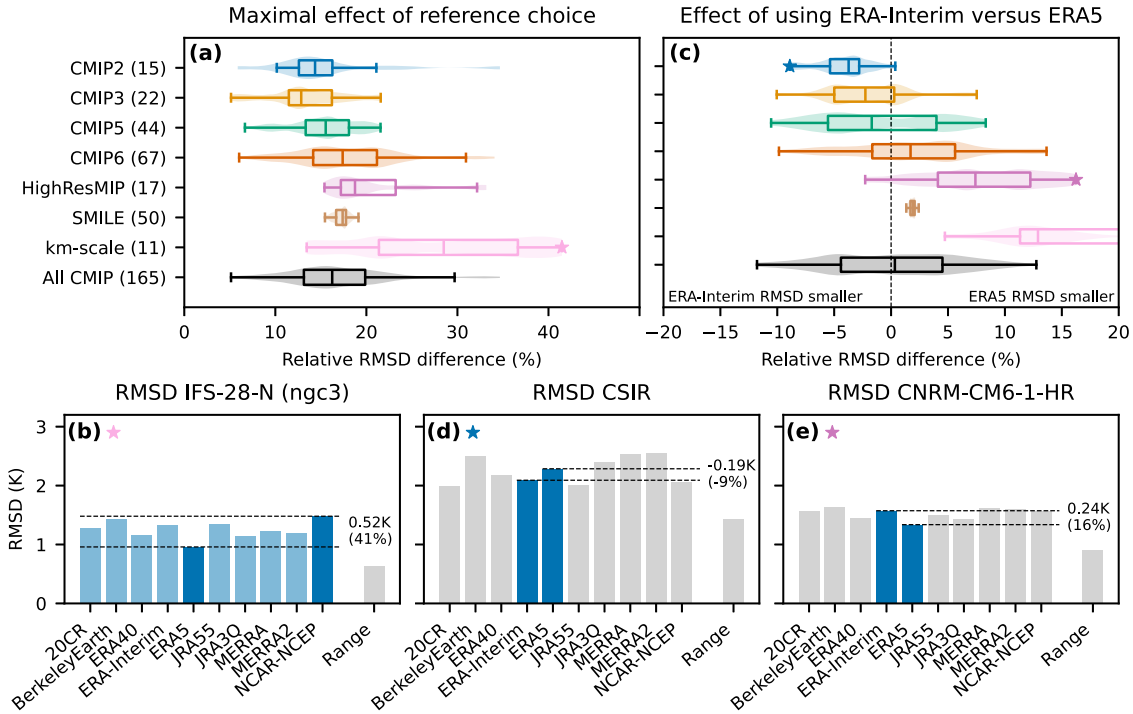

Fig. 2 | Impact of the reference choice on model performance. a Difference between the maximum and minimum RMSD divided by the mean RMSD. The RMSD is calculated for each model using each of the 10 different reference datasets, in turn, as showcased in (**b**) for IFS-28-N (ngc3). **c** Difference in RMSD when using one of two references: ERA-Interim or its successor ERA5, as showcased for **d** CSIR and **e** CNRM-CM6-1-HR. **b**, **d**, **e** The horizontal dashed lines and corresponding numbers give the (absolute and relative) difference between the indicated references.

being quite far away from several other references (Berkeley Earth and NCAR-NCEP) leading to an effect of the reference choice on RMSD of more than 0.5 K or 41%. This example also shows that older or superseded references are not (necessarily) dominating the effect of the reference choice: several of the older references (such as ERA40, ERA-Interim, MERRA) rather lead to intermediate RMSD values (Fig. 2b).

The effect of reference choice also persists when comparing only two of the most frequently used references: ERA-Interim and its successor ERA5. Simply switching from ERA-Interim to ERA5 as a reference (which is what silently happened in the literature after the advent of ERA5 in the mid-2010s) can lead to differences in model performance (Fig. 2c). More importantly, the effect of the reference choice on model performance is systematic across model generations: almost all CMIP2 models have a lower RMSD when using ERA-Interim as reference (median 4%), while almost all HighResMIP (median 7%) and all km-scale (median 13%) models are closer to ERA5 (Fig. 2c). Figure 2d, e shows examples for the CMIP2 model with the strongest "preference" for ERA-Interim and the HighResMIP model with the strongest "preference" for ERA5. In the supplement, we also show an analysis of the effect for the three modern reanalysis datasets (ERA5, JRA3Q, MERRA2; Fig. S7) and an overview of all models (Fig. S8 and Table S5). These findings have implications for assessments of model performance across generations as they show that the improvement over time partly depends on the choice of reference: using ERA5 leads to a higher performance increase of km-scale models compared to early CMIP models relative to continued usage of ERA-Interim as reference (Fig. 2c).

The influence of model resolution and tuning

Finally, we discuss the role of model resolution for the evolution of model performance in a long-term context, revealing a clear relationship that holds across models and generations (Fig. 3), using the RMSD based on the models' distance to the full reference range (Eq. (1)). This relation is strongest across all 148 CMIP models (excluding HighResMIP, see below) where the RMSD improves by −0.30 K/100 km (95% range: −0.36 K/100 km to −0.23 K/100 km), less but similarly strong within generations (mean across CMIP slopes: −0.25 K/100 km), and weakest for 13 high-low resolution pairs of the same models (mean across pairs: −0.18 K/100 km; see Table S1 for details).

The greater improvement across all 148 CMIP models is not surprising, as the slope clearly also includes other drivers beyond resolution. Most importantly, the fact that higher-resolved models are, on average, also more recent, adds growing knowledge and methodological advancements with time to the slope. Estimating the resolution effect on performance only from models developed at a similar time (i.e., from models within each generation) reduces the effect of time, leading to the estimate of −0.25 K/100 km. Calculating it from pairs of models that were run at two different resolutions could be considered the cleanest estimate, but it is limited to only 13 model pairs (7 of which are from CMIP6) and is, hence, less robust. In addition, a model might be run at two resolutions, but only be fully tuned to one of them.

In fact, the influence of tuning is crucial as the example of HighResMIP shows: the HighResMIP protocol explicitly states that additional tuning of the higher-resolution versions should be avoided[27], so that the pure resolution effect is relatively well isolated. From the six HighResMIP model pairs, five worsen with increasing resolution for our metric (Table S1), and the slope across all 17 HighResMIP models is only −0.08 K/100 km. This is a clear indication that, while resolution increase might be a necessary step for model improvement, it is by no means sufficient, and careful tuning to a given resolution is perhaps even more crucial. This underlines the potential of km-scale models, some of which already perform well despite being hardly tuned at this point.

Summary and discussion

We have shown that the model representation of the 20-year mean 2 m surface air temperature pattern has steadily improved over the last three decades in a multi-model mean view but less so when considering the best models in a generation (Fig. 1). Yet, ICON and IFS, two prototype km-scale global climate models have the potential to outperform the best established

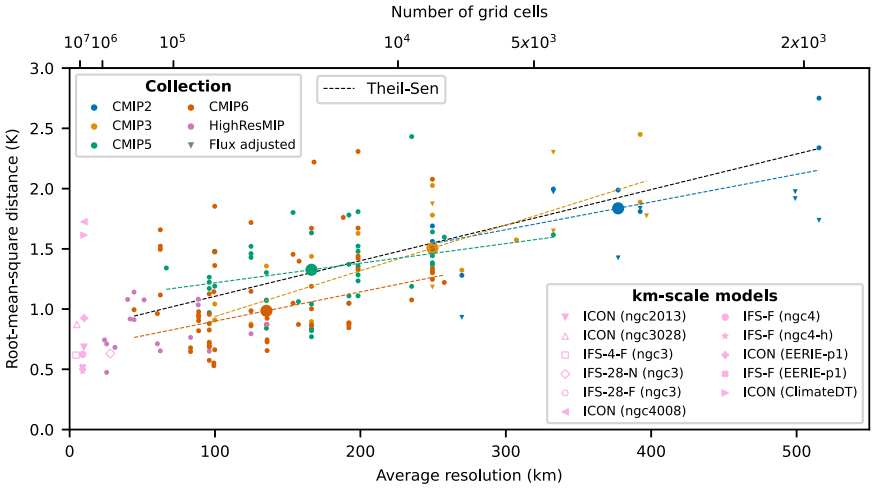

Fig. 3 | Model performance as a function of model resolution. RMSD to the full reference range as a function of model resolution for all 165 CMIP and 11 km-scale models. Unfilled symbols indicate runs with less than 20 years. The dashed lines show Theil-Sen regressions: (black) over all 148 CMIP2,3,5, and 6 models (excluding HighResMIP) and (colored) over all models within each generation. The large dots indicate the median values for each generation.

models from CMIP6 but, at the same time, can perform worse than a median CMIP model, indicating that resolution might be necessary for further model improvement but is not sufficient. Our findings also revealed that increasing model performance leads to an increasing effect of the reference choice, which can exceed 30% based on 10 different observation-based datasets developed in the last decades (Fig. 2). Even when restricting the choice to only modern reanalysis products, systematic differences exceeding 15% still exist (Fig. S7). Crucially, we note that while focusing solely on surface temperature has enabled the long-term multi-model, multi-reference perspective we provide, it is also a clear limitation of our work as it omits many climate system features important for a comprehensive model evaluation.

Finally, we also provided an alternative view on model performance as a function of the models' spatial resolution instead of the more established generational view to investigate the resolution dependence of model performance (Fig. 3) similar, for example, to work by ref. 45. While we found clear signs for an increase in model performance with resolution for well-tuned CMIP models, we also found a decrease in the same relationship for five out of six HighResMIP model pairs, which we attribute to the fact that they are not tuned at high resolution, deteriorating their performance. This, again, underscores that resolution alone is not enough to increase model performance.

Improvements of newly developed km-scale models[43] over their coarser predecessors, hence, depend on realizing the full potential resolution provides. This potential can manifest in many domains, from the model output resolution to the representation of physical processes. For example, recent work by ref. 46 has shown that the finer grid spacing more clearly resolves land surface heterogeneity, particularly at mountain ranges and along coastlines in both ICON and IFS, with implications for temperature extremes and impacts. 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 [refs. 46–49, see also Table S2]. This difference in approaches reflects the fact that the grid spacing of about 10 km puts both models in somewhat of a transition zone, which is slightly too coarse to accurately permit explicit convection (often set to about 2–4 km; e.g., ref. 50) but too fine to allow a robust empirical parameterization (about 25–50 km; ref. 51). While our results for mean temperature show that the partly 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[6] and that land-atmosphere coupling is better represented, affecting (extreme) temperatures[52]. Ongoing efforts to run climatological time-periods at 5 km grid spacing globally[53], contributing to push km-scale models towards the explicit convection domain, hence, further contribute to materializing their potential.

As models keep improving and resolution keeps increasing, finding appropriate reference datasets will get more difficult and more important at the same time, as the choice of reference can start to dominate evaluation. The realization of envisioned products like Climate Digital Twins[53] and Earth Virtualization Engines[54] will continue to increase the integration of observation- and model-based information, further complicating evaluation with independent datasets. At the same time, advances in modeling can even lead to a reversal of the evaluation chain. Already, there are examples where models pointed to deficits in observation-based products[55,56]. Ultimately, we recommend including reference uncertainty in model evaluations whenever possible to document model-reference dependencies arising from the development cycle, and to account for these dependencies in model evaluations. This holds particularly when tracking climate model evolution over time and when evaluating the latest km-scale and upcoming CMIP7 models.

Data and methods

Model data

We use all available CMIP models that provide 2 m surface air temperature from pre-industrial control (piControl) simulations. In addition, we also use all models from the CMIP6-generation high-resolution model intercomparison project (HighResMIP), which provide the same variable for the 1950-control simulations. This leads to a total of 165 models detailed in Table S5. In our analysis, we treat all models and model versions equally in what has been termed "model democracy" and do not account for any model dependencies[57,58]. In addition, we use 50 members from MPI-ESM1-2-LR in the historical 1980–1999 period to quantify internal variability. The same model, driven by pre-industrial forcing, is also included in CMIP6, and we use the direct comparison to show the effect of choosing piControl over the historical period in the rest of the models. Finally, we use two km-scale models (ICON and IFS) in a total of 11 different configurations from the nextGEMS (https://nextgems-h2020.eu/), EERIE, and DestinE (https://destine.ecmwf.int/) projects (see Table S2 for a full list).

All spatial fields are conservatively regridded to 2.5° × 2.5°, which means coarsening the native resolution for almost all models beyond CMIP2. For some of the km-scale models, conservative regridding is not possible and, therefore, the different regridding strategies are detailed in Table S3.

Additional details on the km-scale models

The km-scale models used in this study are set apart from the CMIP models by a considerably higher resolution, using a grid spacing of about 10 km or

even finer, both in the atmosphere and ocean[43]. Compared to conventional CMIP models, the finer grid spacing permits atmospheric deep convection and (sub)mesoscale eddies in the ocean, which allows for turning off mesoscale eddy parameterizations[59,60]. Also, the heterogeneity of the land surface is better represented, and does include cities or urban areas in the case of the IFS model[5,46]. Several of the km-scale model versions used here are (superseded) prototypes or still in development. They are not fully tuned, and that might still include various bugs. Since only limited documentation of these models exists in the literature to date, we summarize their main features in Table S2 and below.

All the ICON simulations use the ICON model run in its Sapphire configuration described in ref. 4. The version identifiers ngc2(013), ngc3(028), and ngc4(008) correspond to different model development cycles of nextGEMS and include various model updates and bug fixes. Compared to the version described in ref. 4, ICON (ngc2013) includes river discharge, uses the RTE-RRTMGP radiation scheme instead of PSRAD, the TTE scheme instead of Smagorinsky as turbulence scheme in the atmosphere, and employs z-star coordinates as well as some updated parameter values in the ocean. ICON (ngc3028)[61] employs the Smagorinsky scheme again as a turbulence scheme in the atmosphere, includes energy fixes in the microphysics, bug fixes in the computation of the surface fluxes, and new land external parameters. In ICON (ngc4008)[62], the turbulence interface has been rewritten, energy leaks have been fixed, and the cloud amount has been retuned by adding a stability correction in the turbulence scheme and adapting the cloud inhomogeneity factor in the radiation. The number of vertical levels in the ocean has decreased, but thin layers are kept in the upper ocean. The ICON (ngc4008) exhibits a stable climate, with no strong cooling of the temperature as documented in ref. 43, and no energy leak. Also, in contrast to the previous ICON versions, which use fixed greenhouse gas and aerosol forcing (set to 2020 conditions), ICON (ngc4008) is integrated using the SSP3-7.0 scenario. The global mean warming between the first and last 10 years of the 30-year integration is about 0.7 K. The ICON EERIE-p1 (phase 1) simulation employs the same radiation and atmospheric turbulence schemes as ICON (ngc2013), as it provided a stable climate, and uses the z-star coordinates and 72 vertical levels for the ocean, like ICON (ngc4008). The ICON version from the Destination Earth project, finally, is labeled ClimateDT as it serves as a basis to build a Climate Digital Twin, as is described in ref. 53.

The IFS simulations used in this manuscript were performed with two different ocean models, namely with the IFS-FESOM (IFS-F) and IFS-NEMO (IFS-N) model configurations as described in detail by ref. 5. While the 28 km simulation (IFS-28-F) with a 0.25° ocean uses the Gent-McWilliams (GM) parameterization for mesoscale eddies[59,60] in the implementation described by ref. 63, all 5 km ocean eddy-resolving simulations from EERIE and nextGEMS do not use GM. Consistent with the other nextGEMS model ICON, the IFS scenario simulation (IFS-F ngc4) was also performed with the SSP3-7.0 scenario for the period 2020-2049. For consistency within nextGEMS and despite the coarser atmospheric resolution of 9 km in IFS-F ngc4, a reduced cloud-base mass flux has been applied as for the earlier 4.4 km nextGEMS simulation (IFS-4-F ngc3). The global mean surface warming of 1.15 K between the last and the first 10 years of the 30-year integration is higher by 0.45 K than for the ICON model. The IFS-F model configuration for the EERIE simulation (IFS-F EERIE-p1) mirrors that of the IFS-F (ngc4), with the exception of deep convection parameterization, where the reduced cloud-base mass flux is not implemented. Moreover, compared to IFS-F (ngc4), the IFS-F (EERIE-p1) is a simulation with constant 1950 radiative forcing[64].

Observation-based references

As references, we use nine reanalysis datasets, which provide 2 m surface air temperature, and Berkeley Earth, which provides blended 2 m surface air temperature over land and sea ice and sea surface temperature over open ocean. We note that other direct observational temperature datasets with global coverage, such as HadCRUT[65] or GISTEMP[66] do not provide absolute temperature fields but only changes from a base period and can, therefore, not be used here, and other products are not available globally[67,68].

All reference datasets are conservatively regridded to a common 2.5° × 2.5° grid equivalent to the models.

Definition of the model performance metric

We use the 20-year mean of the spatially resolved 2 m surface air temperature field as a basis for evaluating model performance. For the CMIP models, the climatology is calculated from 20 years at the end of the piControl run. For the CMIP6-generation HighResMIP, no pre-industrial control run is available, and we use 1950-control instead. For the SMILE and the reference datasets, we use the 1980–1999 mean from the historical period. The km-scale simulations are based on different scenarios as summarized in Table S2 and we use whatever forcing is available.

We use piControl runs instead of historical runs in CMIP to (1) maximize data availability and (2) maximize model consistency. (1) CMIP2 allows unique insights into several models from the early days of coupled atmosphere-ocean model intercomparison, but it only provides piControl runs since historically forced runs only became part of the CMIP protocol with CMIP3[23]. (2) In our main comparison we prioritize consistency between models from different generations, which is ensured by the use of piControl. The historical forcing, in contrast, has been updated between generations[69].

To correct for the first-order differences due to the different absolute climate states between the datasets, we remove the global mean temperature from each dataset separately. We are, hence, interested in the models' ability to simulate the spatial patterns of temperature, not its absolute value[70]. Correcting for differing global means between piControl and historical runs clearly does not offset all effects of warming, with land warming faster than the ocean and high latitudes faster than lower ones. To estimate the global mean effect of remaining differences, we compare the historical period in the 50-member SMILE with the piControl run from the same model in Fig. 1 of the main manuscript. In addition, we also provide a version of Fig. 1 including only models that provide historical simulations (Fig. S6) as well as the difference between piControl and historical for each model (Fig. S12).

Mathematically formulated, the performance metric τ is calculated from the 3-dimensional temperatures T following

$$\tau'(\text{year}, \text{lat}, \text{lon}) = T(\text{year}, \text{lat}, \text{lon}) - \text{Mean}_{\text{lat,lon}}\left(T(\text{year}, \text{lat}, \text{lon})\right) \quad (2)$$

$$\tau(\text{lat}, \text{lon}) = \text{Mean}_{\text{year}}\left(\tau'(\text{year}, \text{lat}, \text{lon})\right), \quad (3)$$

where the $\text{Mean}_{\text{lat,lon}}$ operator uses weighting based on the cosine of the latitude.

Estimating model performance relative to 10 observation-based references

We calculate model performance according to Eq. (1). For a single reference, the metric is simply the model-reference distance at each grid cell (with $\tau_{\text{ref\_min}} = \tau_{\text{ref\_max}}$), for more than one reference, the distance at a given grid cell is zero if the model falls between the references and increases linearly outside. Examples for the resulting difference fields are given in Fig. 1 in the main manuscript, and a repository of all models can be found online in ref. 38.

To test the robustness of the results to our choice of reference datasets we provide two additional approaches: (1) using only the three latest full reanalyses ERA5, JRA3Q, and MERRA2 as reference (Fig. S4) and (2) fitting a Gaussian to the 10 reference datasets at each grid cell and using the 5th and 95th percentiles as reference range (Fig. S5).

From the distance maps $\Delta\tau$, we calculate the area-weighted spatial RMSD and the area-weighted fraction of grid cells that fall inside the reference range (i.e., have a distance $\Delta\tau$ of zero).

We also calculate the RMSD and fraction inside the reference range of the reference datasets in a leave-one-out cross-validation setting. For each of the 10 observation-based datasets, we use the remaining nine as

references in turn. This cross-validation approach serves as a fair baseline for the model performance and can reveal when models start to outperform the reference datasets.

Calculation of performance-resolution slopes

We approximate mean model resolution Δd for all CMIP models using a pragmatic approach: it is calculated as the square root of the surface of a sphere with the mean radius of the Earth divided by the number of grid cells in the native atmospheric resolution files:

$$\Delta d = \sqrt{\frac{4\pi \times 6371\ \mathrm{km}^2}{\mathrm{nr\_grid\_cells}}}\,[\mathrm{km}] \quad (4)$$

We note that the provided native resolution is not necessarily the real model resolution. However, from the CMIP6 models provided via the Earth System Grid Federation, the vast majority are on the native model grid. In addition, it is reasonable to assume that even for models provided on different grids, the effective resolution does not change much from the native model grid, so that our approximation holds in general.

For the reference datasets and the km-scale models, we use values provided in the respective documentation.

To calculate changes of model performance with resolution, we use the Theil-Sen estimator as implemented in Python's SciPy package. It computes the slope as the median over all possible pairs of models with differing resolution values (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.theilslopes.html).

For the case of high-low resolution model pairs, we compute the median only over pairs from the same model.

Data availability

The processed climatologies used for this study are freely available from ref. 38, the raw model output is available upon reasonable request.

Code availability

The code used to process and visualize the main results is available from ref. 38.

Received: 26 September 2025; Accepted: 30 March 2026;
Published online: 16 April 2026

References

1. Manabe, S. & Bryan, K. Climate calculations with a combined ocean-atmosphere model. *J. Atmos. Sci.* **26**, 786–789 (1969).
2. Meehl, G. A., Boer, G. J., Covey, C., Latif, M. & Stouffer, R. J. The Coupled Model Intercomparison Project (CMIP). *Bull. Am. Meteorol. Soc.* **81**, 313–318 (2000).
3. Flato, G. M. Earth system models: an overview. *WIREs Clim. Change* **2**, 783–800 (2011).
4. Hohenegger, C. et al. ICON-Sapphire: Simulating the components of the Earth system and their interactions at kilometer and subkilometer scales. *Geosci. Model Dev.* **16**, 779–811 (2023).
5. Rackow, T. et al. Multi-year simulations at kilometre scale with the Integrated Forecasting System coupled to FESOM2.5 and NEMOv3.4. *Geosci. Model Dev.* **18**, 33–69 (2025).
6. Takasuka, D. et al. How can we improve the seamless representation of climatological statistics and weather toward reliable global K-scale climate simulations? *J. Adv. Model. Earth Syst.* **16**, e2023MS003701 (2024).
7. Schneider, T., Leung, L. R. & Wills, R. C. J. Opinion: Optimizing climate models with process knowledge, resolution, and artificial intelligence. *Atmos. Chem. Phys.* **24**, 7041–7062 (2024).
8. Schär, C. et al. Kilometer-scale climate models: prospects and challenges. *Bull. Am. Meteorol. Soc.* **101**, E567–E587 (2019).
9. McGuffie, K. & Henderson-Sellers, A. Forty years of numerical climate modelling. *Int. J. Climatol.* **21**, 1067–1109 (2001).
10. Edwards, P. N. History of climate modeling. *Wiley Interdiscip. Rev. Clim. Change* **2**, 128–139 (2011).
11. Treut, H. L. et al. Historical overview of climate change science. In (eds Solomon, S. et al.) *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, 2007).
12. Duffy, P. B., Bell, J., Covey, C. & Sloan, L. Effect of flux adjustments on temperature variability in climate models. *Geophys. Res. Lett.* **27**, 763–766 (2000).
13. Bordoni, S., Kang, S. M., Shaw, T. A., Simpson, I. R. & Zanna, L. The futures of climate modeling. *npj Clim. Atmos. Sci.* **8**, 1–6 (2025).
14. Tokarska, K. B. et al. Past warming trend constrains future warming in CMIP6 models. *Sci. Adv.* **6**, eaaz9549 (2020).
15. Brunner, L. et al. Reduced global warming from CMIP6 projections when weighting models by performance and independence. *Earth Syst. Dyn.* **11**, 995–1012 (2020).
16. Palmer, T. E. et al. Performance-based sub-selection of CMIP6 models for impact assessments in Europe. *Earth Syst. Dyn.* **14**, 457–483 (2023).
17. Meehl, G. A., Boer, G. J., Covey, C., Latif, M. & Stouffer, R. J. Intercomparison makes for a better climate model. *Eos*. **78**, 445–451 (1997).
18. Bock, L. et al. Quantifying progress across different CMIP phases with the ESMValTool. *J. Geophys. Res. Atmos.* **125**, 1–28 (2020).
19. IPCC. 2021: summary for policymakers. in *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Masson-Delmotte, V. et al.) (IPCC, 2021).
20. Carvalho, D. et al. How well have CMIP3, CMIP5 and CMIP6 future climate projections portrayed the recently observed warming. *Sci. Rep.* **12**, 11983 (2022).
21. Reichler, T. & Kim, J. How well do coupled models simulate today' s climate? *Bull. Am. Meteorol. Soc.* **819**, 303–311 (2008).
22. Covey, C. et al. An overview of results from the Coupled Model Intercomparison Project. *Glob. Planet. Change* **37**, 103–133 (2003).
23. Meehl, G. A., Covey, C., McAvaney, B., Latif, M. & Stouffer, R. J. Overview of the coupled model intercomparison project. *Bull. Am. Meteorol. Soc.* **86**, 89–93 (2005).
24. Meehl, G. A. et al. The WCRP CMIP3 multimodel dataset: a new era in climatic change research. *Bull. Am. Meteorol. Soc.* **88**, 1383–1394 (2007).
25. Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93**, 485–498 (2012).
26. Eyring, V. et al. Earth System Model Evaluation Tool (ESMValTool) v2.0 – an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP. *Geosci. Model Dev.* **13**, 3383–3438 (2020).
27. Haarsma, R. J. et al. High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6. *Geosci. Model Dev.* **9**, 4185–4208 (2016).
28. Slivinski, L. C. et al. Towards a more reliable historical reanalysis: improvements for version 3 of the Twentieth Century Reanalysis system. *Q. J. R. Meteorol. Soc.* **145**, 2876–2908 (2019).
29. Rohde, R. A. & Hausfather, Z. The Berkeley Earth Land/Ocean Temperature Record. *Earth Syst. Sci. Data* **12**, 3469–3479 (2020).
30. Uppala, S. M. et al. The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131**, 2961–3012 (2005).
31. Dee, D. P. et al. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137**, 553–597 (2011).

32. Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
33. Kobayashi, S. et al. The JRA-55 Reanalysis: general specifications and basic characteristics. *J. Meteorol. Soc. Jpn. Ser. II* **93**, 5–48 (2015).
34. Kosaka, Y. et al. The JRA-3Q Reanalysis. *J. Meteorol. Soc. Jpn. Ser. II* **102**, 49–109 (2024).
35. Rienecker, M. M. et al. MERRA: NASA's modern-era retrospective analysis for research and applications. *J. Clim.* **24**, 3624–3648 (2011).
36. Gelaro, R. et al. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *J. Clim.* **30**, 5419–5454 (2017).
37. Kalnay, E. et al. The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77**, 437–471 (1996).
38. Brunner, L. Supplement for: three decades of simulating global temperature patterns with coupled global climate models. https://doi.org/10.5281/zenodo.18853482 (2026).
39. Davy, R. & Outten, S. The Arctic Surface Climate in CMIP6: Status and Developments since CMIP5. *J. Climate*, **33**, 8047–8068 (2020).
40. Luo, F., Ying, J., Liu, T. & Chen, D. Origins of Southern Ocean warm sea surface temperature bias in CMIP6 models. *npj Clim. Atmos. Sci.* **6**, 127 (2023).
41. Olonscheck, D. et al. The New Max Planck Institute Grand Ensemble With CMIP6 Forcing and High-Frequency Model Output. *J. Adv. Model. Earth Syst.* **15**, e2023MS003790 (2023).
42. Wedi, N. P. Increasing horizontal resolution in numerical weather prediction and climate simulations: Illusion or panacea? *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **372**, 20130289 (2014).
43. Segura, H. et al. nextGEMS: entering the era of kilometer-scale Earth system modeling. *Geosci. Model Dev.* **18**, 7735–7761 (2025).
44. Zumwald, M. et al. Understanding and assessing uncertainty of observational climate datasets for model evaluation using ensembles. *WIREs Clim. Change* **11**, e654 (2020).
45. Masson, D. & Knutti, R. Spatial-scale dependence of climate model performance in the CMIP3 ensemble. *J. Clim.* **24**, 2680–2692 (2011).
46. Brunner, L. et al. A global perspective on the spatial representation of climate extremes from km-scale models. *Environ. Res. Lett.* **20**, 074054 (2025).
47. Spät, D., Biasutti, M., Schuhbauer, D. & Voigt, A. Autocorrelation—a simple diagnostic for tropical precipitation variability in global kilometer-scale climate models. *Geophys. Res. Lett.* **51**, e2024GL108856 (2024).
48. Wille, J. D., Koch, R., Becker, T. & Fischer, E. Extreme precipitation depiction in convection-permitting Earth system models within the nextGEMS Project. *J. Adv. Model. Earth Syst.* **17**, e2024MS004840 (2025).
49. Takasuka, D., Becker, T. & Bao, J. Precipitation characteristics and thermodynamic-convection coupling in global kilometer-scale simulations. *J. Adv. Model. Earth Syst.* **18**, e2025MS005343 https://doi.org/10.1029/2025MS005343 (2026).
50. Prein, A. F., Rasmussen, R. M., Wang, D. & Giangrande, S. E. Sensitivity of organized convective storms to model grid spacing in current and future climates. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **379**, 20190546 (2021).
51. Vergara-Temprado, J., Ban, N., Panosetti, D., Schlemmer, L. & Schär, C. Climate models permit convection at much coarser resolutions than previously considered. *J. Clim.* **33**, 1915–1933 (2020).
52. Lee, J. et al. Systematic and objective evaluation of Earth system models: PCMDI Metrics Package (PMP) version 3. *Geosci. Model Dev.* **17**, 3919–3948 (2024).
53. Doblas-Reyes, F. J. et al. *The Destination Earth Digital Twin for Climate Change Adaptation* 1–41 (EGUsphere, 2025).
54. Stevens, B. et al. Earth Virtualization Engines (EVE). *Earth Syst. Sci. Data* **16**, 2113–2122 (2024).
55. Sippel, S. et al. Early-twentieth-century cold bias in ocean surface temperature observations. *Nature* **635**, 618–624 (2024).
56. Medhaug, I., Stolpe, M. B., Fischer, E. M. & Knutti, R. Reconciling controversies about the 'global warming hiatus'. *Nature* **545**, 41–47 (2017).
57. Knutti, R. The end of model democracy? *Clim. Change* **102**, 395–404 (2010).
58. Merrifield, A. L., Brunner, L., Lorenz, R., Humphrey, V. & Knutti, R. Climate model Selection by Independence, Performance, and Spread (ClimSIPS v1.0.1) for regional applications. *Geosci. Model Dev.* **16**, 4715–4747 (2023).
59. Gent, P. R. & Mcwilliams, J. C. IsopycnaL MIXING IN OCEAN CIRCULATION MODels. *J. Phys. Oceanogr.* **20**, 150–155 (1990).
60. Gent, P. R., Willebrand, J., McDougall, T. J. & McWilliams, J. C. Parameterizing Eddy-induced tracer transports in ocean circulation models. *J. Phys. Oceanogr.* **25**, 463–474 (1995).
61. Koldunov, N. et al. nextgems: output of the model development cycle 3 simulations for ICON and IFS. https://doi.org/10.26050/WDCC/nextGEMS_cyc3 (2023).
62. Wieners, K.-H. et al. nextgems: output of the production simulations for ICON and IFS https://www.wdc-climate.de/ui/entry?acronym=nextGEMS_prod (2024).
63. Scholz, P. et al. Assessment of the Finite-volumE Sea ice-Ocean Model (FESOM2.0)—Part 1: description of selected key model elements and comparison to its predecessor version. *Geosci. Model Dev.* **12**, 4875–4899 (2019).
64. Ghosh, R. et al. EERIE: IFS-FESOM TC01279/NG5 (9km, 5km): spinup-1950. https://www.wdc-climate.de/ui/entry?acronym=DKRZ_LTA_1344_dsg0001 (2024).
65. Morice, C. P. et al. An updated assessment of near-surface temperature change from 1850: the HadCRUT5 Data Set. *J. Geophys. Res. Atmos.* **126**, e2019JD032361 (2021).
66. Lenssen, N. J. L. et al. Improvements in the GISTEMP Uncertainty Model. *J. Geophys. Res. Atmos.* **124**, 6307–6326 (2019).
67. Chan, D., Gebbie, G., Huybers, P. & Kent, E. C. A Dynamically Consistent ENsemble of Temperature at the Earth surface since 1850 from the DCENT dataset. *Sci. Data* **11**, 953 (2024).
68. Huang, B. et al. Improvements of the Daily Optimum Interpolation Sea Surface Temperature (DOISST) Version 2.1. *J. Clim.* **34**, 2923–2939 (2021).
69. Meinshausen, M. et al. Historical greenhouse gas concentrations for climate modelling (CMIP6). *Geosci. Model Dev.* **10**, 2057–2116 (2017).
70. Mauritsen, T. et al. Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and its response to increasing CO_2. *J. Adv. Model. Earth Syst.* **11**, 998–1038 (2019).

Acknowledgements

The authors thank Emanuel Dutra, Benjamin Poschlod, and Adrian Tompkins for helpful discussions and Benjamin Blanz, Blaž Gasparini, Dirk Olonscheck, Ulrike Proske, Irina Sandu, Josipa Zupanic, and three anonymous reviewers from an earlier submission for their comments on the manuscript. The authors thank all the data providers for making their data available (see Tables S9 and S10 for detailed acknowledgments). The work presented in this paper has been produced in the context of the European Union's Destination Earth Initiative and relates to tasks entrusted by the European Union to the European Centre for Medium-Range Weather Forecasts, implementing part of this Initiative with funding by the European Union. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2037 "CLICCS—Climate, Climatic Change, and Society"—Project Number: 390683824. This research has also been supported by the European Commission Horizon 2020 Framework

1008 Programme nextGEMS (grant no. 101003470) and EERIE (grant no. 101081383). The EERIE IFS-FESOM coupled spin-up simulation was performed using high-performance computing resources provided by the Jülich Supercomputing Centre (JSC) on the JUWELS system as well as the Levante supercomputer at the Deutsches Klimarechenzentrum (DKRZ). Data storage, technical support, and data hosting were provided by DKRZ, including dissemination via the World Data Center for Climate (WDCC). We acknowledge financial support from the Open Access Publication Fund of University of Hamburg.

Author contributions

Conceptualization, methodology, investigation, visualization, and project administration by L.B. Original draft written by L.B. supported by A.V. C.H., R.G., D.P., and T.R. provided information and text about the km-scale models, L.H. provided information and text about the reference datasets, and R.K. contributed to the discussion. Original draft review and editing by all authors.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43247-026-03497-w.

Correspondence and requests for materials should be addressed to Lukas Brunner.

Peer review information *Communications Earth and Environment* thanks Qing Bao and Dan Fu for their contribution to the peer review of this work. Primary Handling Editor: ChenRui Diao. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026