



# OPEN Reduced cloud cover errors in a hybrid AI-climate model through equation discovery and automatic tuning

Arthur Grundner<sup>1✉</sup>, Tom Beuchler<sup>2,3</sup>, Julien Savre<sup>1</sup>, Axel Lauer<sup>1</sup>, Manuel Schlund<sup>1</sup> & Veronika Eyring<sup>1,4</sup>

Cloud-related parameterizations remain a leading source of uncertainty in climate projections. Although machine learning holds promise for Earth system models (ESMs), many data-driven parameterizations lack interpretability, physical consistency, and smooth integration into ESMs. Here, a two-step method is presented to improve a climate model with data-driven parameterizations. First, we incorporate a physically consistent cloud cover parameterization—derived from storm-resolving simulations via symbolic regression, preserving interpretability while enhancing accuracy—into the ICON global atmospheric model. Second, we apply the gradient-free Nelder–Mead optimizer to automatically recalibrate the hybrid model against Earth observations, tuning in nested stages (2-, 7-, 30- and 365-day runs) to ensure stability and tractability. The tuned hybrid model substantially reduces long-standing biases in cloud cover—particularly over the Southern Ocean (by 75%) and subtropical stratocumulus regions (by 44%)—and remains robust under +4K surface warming. These results demonstrate that interpretable machine-learned parameterizations, paired with practical tuning, can efficiently and transparently strengthen ESM fidelity.

When numerically modeling the dynamics of a physical field, it needs to be overlaid by a discrete grid. Processes that are happening inside individual grid cells and thus cannot be resolved may have a significant impact on the overall dynamics and must therefore be inferred from the available grid-scale information. In climate models, so-called parameterizations tackle this task<sup>1</sup>. It is natural to learn from data where (e.g., process) knowledge is lacking, therefore, the development of data-driven and machine learning based parameterizations has become widespread in recent years<sup>2–7</sup>. Because of their representation power, neural networks are a common choice for data-driven modeling<sup>8</sup>, despite coming with some clear and significant drawbacks: They often deteriorate the interpretability of climate models, making it difficult to demonstrate their physical consistency, they degrade the climate model's applicability to climate change scenarios<sup>9</sup>, and they increase the computational burden when replacing a parameterization that used to be formulated as a simple equation<sup>10</sup>.

In the development of all types of data-driven parameterizations, the training and testing of a scheme is performed either 'offline' (i.e., in a stand-alone fashion on a reference dataset) or 'online' (i.e., implemented in and interacting with the dynamics of a climate model)<sup>5,11</sup>. Online training requires either a differentiable host climate model<sup>12</sup> or costly inverse methods<sup>13</sup>, making offline training the dominant approach for data-driven parameterizations. However, these offline-trained parameterizations have yet to demonstrate consistent performance online, due to the need for a tuning step to adjust the climate model to its new parameterization<sup>14,15</sup>. Such a tuning step usually requires substantial time and expert knowledge. Its execution also becomes less straightforward as deep learning-based parameterizations no longer include such physics-related tunable parameters. Thus, many of the offline-trained models are never tested online, leading to increased doubts about their usefulness for climate modeling<sup>12</sup>. After all, it is not obvious that a data-driven parameterization can overcome this 'optimization dichotomy'<sup>16</sup> of being trained in one setting while expected to perform well in a different setting.

In this paper we do not only demonstrate the strong online skill of our offline-trained, data-driven parameterization, but also introduce a new automatic tuning pipeline that enables these results. More specifically,

<sup>1</sup>Institut für Physik der Atmosphäre, Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), Oberpfaffenhofen, Germany. <sup>2</sup>Faculty of Geosciences and Environment, University of Lausanne, Lausanne, Switzerland. <sup>3</sup>Expertise Center for Climate Extremes, University of Lausanne, Lausanne, Switzerland. <sup>4</sup>Institute of Environmental Physics (IUP), University of Bremen, Bremen, Germany. ✉email: arthur.grundner@dlr.de

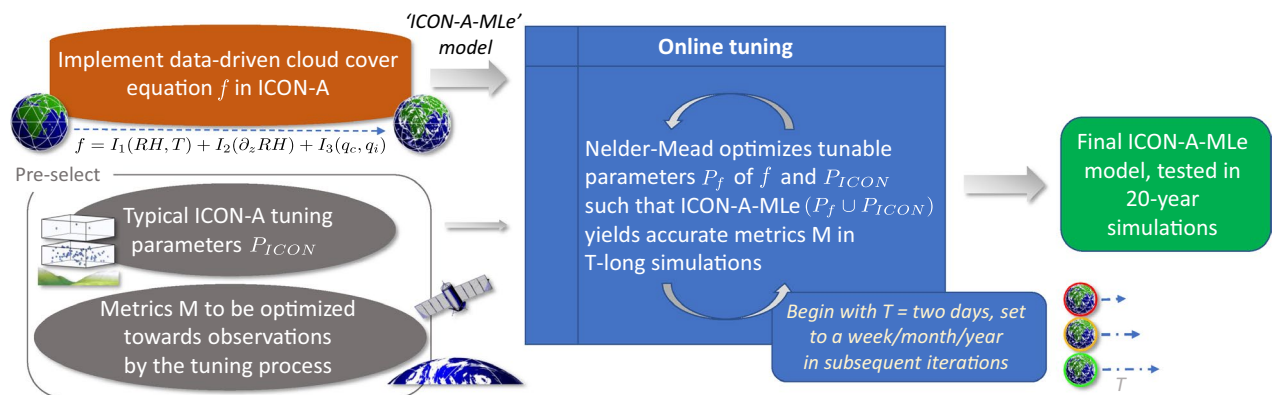
we consider an *analytical equation* (Eq. 3) from<sup>17</sup> that has been trained offline to parameterize cloud fraction, reducing related systematic model errors towards improved climate projections. The equation was derived from a high-resolution simulation using a symbolic regression library<sup>18</sup> for equation discovery<sup>19,20</sup>. It is the product of a data-driven approach that provides accuracy and yet retains the interpretability, physical consistency and computational efficiency of the traditional parameterization. However, its application in a climate model has not yet been demonstrated. To do so, we first implement the data-driven cloud cover equation into ICON-A<sup>21</sup>, the atmospheric component of the ICOSahedral Non-hydrostatic (ICON) climate model. We then set up an automatic tuning procedure following a multi-objective approach<sup>22</sup> to tune the model, hereafter called ‘ICON-A-MLe’ where MLe stands for Machine Learning enhanced. The basic idea is to tune as much as possible using simulations that are as short as possible. Here, we build on the finding that biases resulting from fast physical processes can already be identified in short simulations of a day, week or month<sup>23</sup>. Compared to other recent automated tuning methods for climate models (e.g.,<sup>24,25</sup>), our tuning procedure is simpler and thus more efficient and better generalizable.

We begin by outlining the tuning pipeline and its impact on ICON model results. We then compare the tuned ICON-A-MLe model with observations, as well as to models participating in the Coupled Model Intercomparison Project Phase 6 (CMIP6<sup>26</sup>) and to an automatically and manually tuned version of ICON-A. For the evaluation, we perform a historical Atmospheric Model Intercomparison Project (AMIP) simulation from 1979–1999, which has prescribed sea surface temperatures (SSTs) and sea ice concentrations, and evaluate it with Earth observations to assess its accuracy. For this we use the Earth System Model Evaluation Tool (ESMValTool), an established community diagnostic and performance metrics tool for the evaluation and analysis of Earth system models<sup>27–33</sup>, recently extended to be able to process ICON output without model postprocessing<sup>34</sup>. To isolate the impact of the new scheme from the tuning, we also present a “scheme-swap” experiment with subsequent retuning of the cloud cover schemes. Following this, we provide a physical analysis linking the new scheme’s analytical terms to the improved cloud representation. Finally, we demonstrate its robustness by applying ICON-A-MLe to a significantly warmer climate.

## Overview of the automatic tuning pipeline

To effectively integrate the data-driven cloud cover equation (Eq. 3) into ICON-A, we developed a simple and efficient automatic tuning pipeline (Fig. 1). This pipeline uses the Nelder–Mead optimization algorithm to calibrate a set of 24 physical and empirical parameters—spanning cloud microphysics, convection, and the new parameterization (Table S3)—against observationally derived target metrics (“Methods”).

Utilizing computational resources carefully, the tuning process proceeds in stages, beginning with short 2-day simulations and increasing in duration to 7-day, 30-day, and finally 365-day simulations. At each stage, simulation outputs are compared to observational constraints on key global metrics such as the top-of-the-atmosphere (TOA) radiation balance, shortwave (SW)/longwave (LW) cloud radiative effects (CRE), precipitation, and cloud water paths. The objective function guiding optimization quantifies the normalized distance between simulated values and observational ranges. The shorter the simulations in a given stage, the more those ranges are relaxed informed by the typical variability of a given metric in a reference ICON simulation. Each optimization stage refines the parameter space based on the best result from the previous stage. The final step includes an extrapolated parameter update to balance short- and long-term simulation objectives,



**Fig. 1.** The automatic tuning pipeline of ICON-A-MLe, i.e., the ICON atmospheric climate model including our data-driven cloud cover equation (Eq. 3). The only manual step involves selecting the climate metrics to be optimized, such as globally averaged cloud cover, and determining which tunable parameters should be adjusted. Reasonable target ranges for these metrics are derived from observational and reanalysis data. The core of the tuning pipeline consists of the Nelder–Mead algorithm iteratively searching for a setting of the tunable parameters that enables ICON-A-MLe simulations to optimize all specified metrics. The length of these simulations and the strictness of evaluating the metrics increases with the number of iterations. The underlying strategy leverages the fact that some climate metrics exhibit rapid responses to parameter changes, allowing for efficient feedback and early adjustments. Finally a historical AMIP simulation is benchmarked to Earth observations with the ESMValTool.

followed by a small number of year-long simulations to identify an optimal parameter set. To minimize the objective function within each stage, we use the Nelder-Mead algorithm<sup>35</sup>, a robust, derivative-free method that moves and tumbles a “simplex” (which is a triangle in a 2-dimensional parameter space, a tetrahedron in 3D, and so on) through the parameter space, iteratively discarding the worst-performing vertices. We apply it in this study to automatically find the set of model parameters that minimizes the deviation between our simulation output and observational data. This iterative strategy allows for robust and efficient convergence toward physically realistic climate simulations.

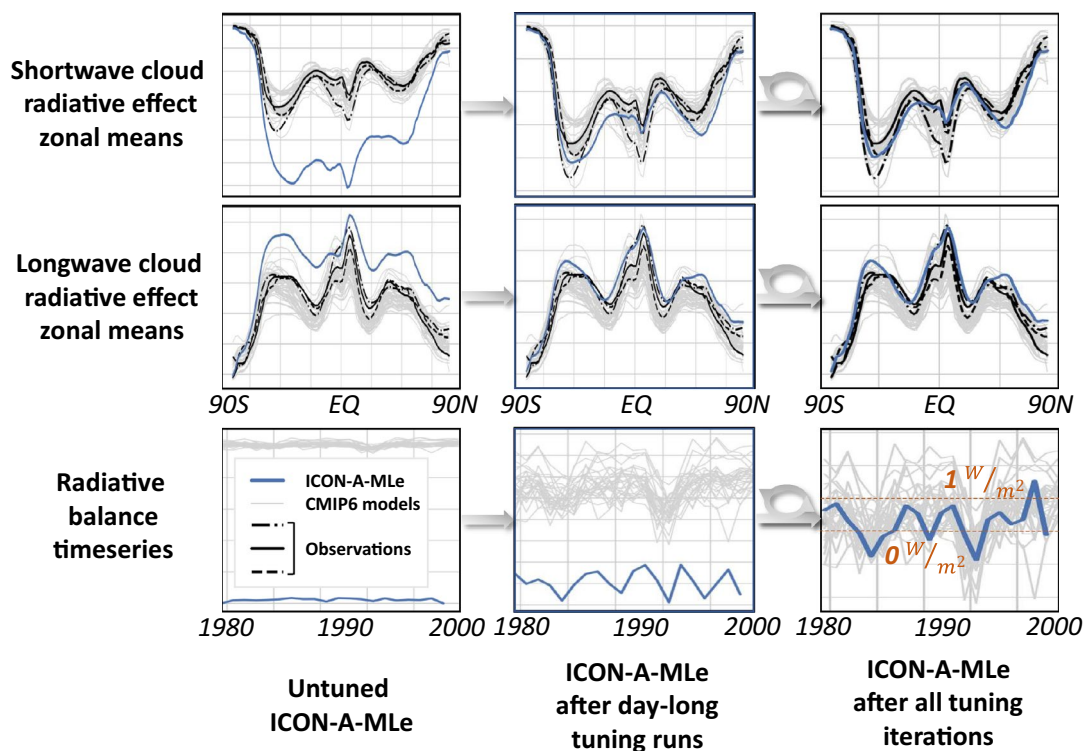
Figure 2 illustrates the impact of the automatic tuning pipeline on the performance of ICON-A-MLe in 20-year simulations for three key radiative metrics (zonal means of the absorbed SW and outgoing LW radiation are provided in Fig. S6). When the data-driven equation (Eq. 3) with its original parameters (Table S3) is first implemented in ICON, these metrics deviate significantly from observations without tuning. After each optimization stage, these metrics gradually improve with the top-of-the-atmosphere radiative balance assuming realistic values at the end of the tuning process.

### Evaluation of 20-year ICON simulations

In this section, we compare the tuned ICON-A-MLe and ICON-A models with observational reference datasets, focusing on the mean state of the simulated climate. Specifically, we conduct a quantitative evaluation of the 20-year historical AMIP simulations referenced in the green box of Fig. 1.

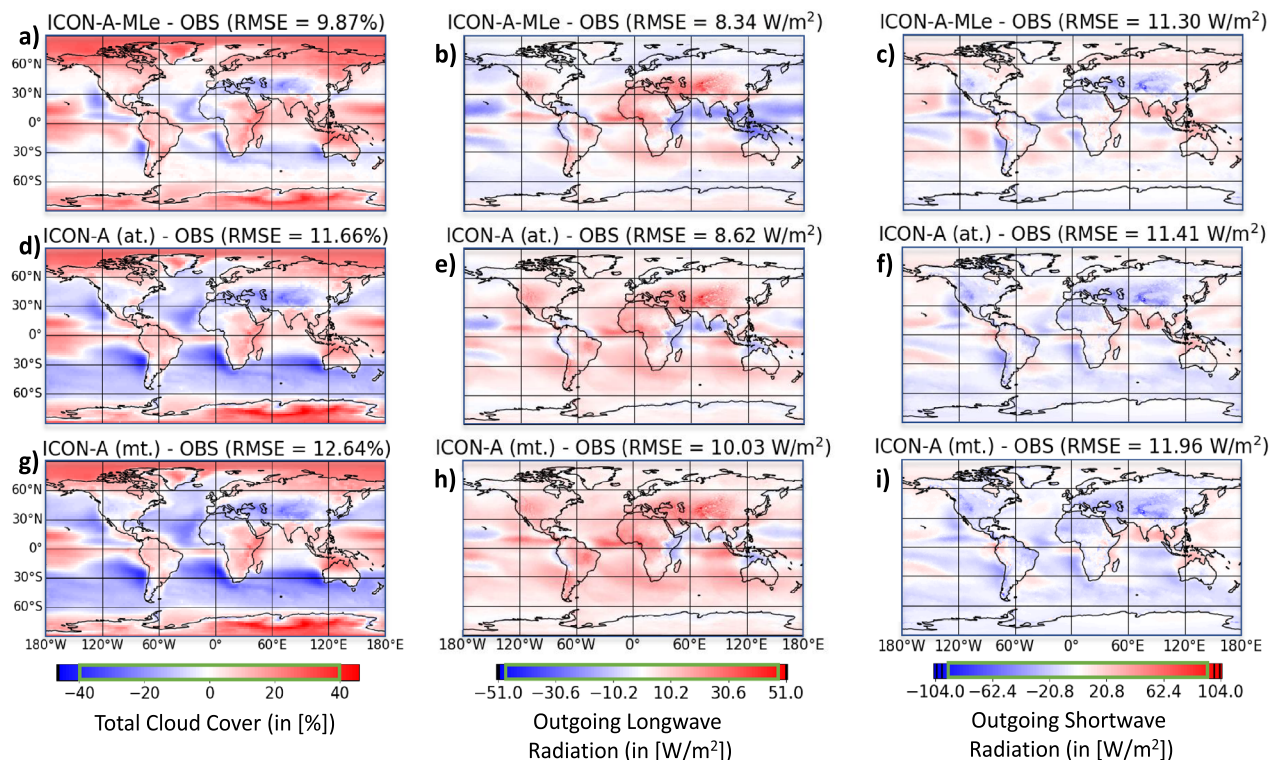
Total cloud cover is the first metric to evaluate when testing a new cloud cover parameterization. Its importance cannot be overstated: clouds are fundamental to the Earth’s energy budget, as they can either cool the Earth by reflecting sunlight or warm it by trapping heat. Their role in modulating precipitation and humidity further underscores their influence on local and regional climates<sup>45</sup>.

The first column of Fig. 3 demonstrates that ICON-A-MLe achieves a notably more accurate representation of the geographical distribution of temporally averaged total cloud cover compared to the manually and automatically tuned ICON-A setups (global RMSE reduced by 15.4% from Panel (d) to (a)). These improvements are most evident in maritime regions, particularly in regions dominated by marine low-level clouds<sup>46</sup>; the Southern Ocean (RMSE reduced by 75.8%), the west coasts of Chile/Peru (by 49.9%), Namibia/Angola (by 54.2%), California (by 24.1%), Morocco (by 41.8%), and Australia (by 53.3%). These are precisely the regions that exhibit the largest positive differences between Panels (a) and (d) (see Fig. S7a). Polewards of 60° N and 60° S, ICON-A significantly overestimates cloud cover. While ICON-A-MLe shows an increased cloudiness over Canada and Greenland even further (locally by up to 23%) compared to ICON-A, it displays reduced cloud



**Fig. 2.** A qualitative evaluation of 20-year ICON-A-MLe simulations using parameter settings extracted at three different stages of the tuning pipeline. The panels display radiative measures computed at the top of the atmosphere, illustrating the model’s progression through the tuning process. Circular arrows denote intermediate tuning steps involving week- and month-long ICON-A-MLe simulations. Observational references are derived from MERRA2<sup>36</sup>, CERES<sup>37</sup>, and ISCCP<sup>38</sup>. The solid gray lines represent historical CMIP6 model simulations<sup>26</sup>, providing a benchmark for comparison.





**Fig. 3.** Biases in three key climate metrics, temporally averaged over 20-year simulations (1979–1999), for our automatically tuned (at.) ICON-A-MLe (first row) and ICON-A models (second row), and a manually tuned (mt.) ICON-A reference (third row). Root-mean-square errors (RMSEs) are computed after remapping the data onto a horizontal grid with nearly identical grid cell sizes. While the ICON-A panels utilize the full range of each colorbar (otherwise indicated by black vertical lines in the colorbars), the values from the ICON-A-MLe model are confined to the green rectangles. To enhance the robustness of our observational reference (OBS), we take an average across multiple observational cloud cover datasets (CLARA-AVHRR<sup>39</sup>, ESA CCI<sup>40</sup>, MODIS<sup>41</sup>, PATMOS<sup>42</sup>), as well as two reanalysis products (MERRA2<sup>36</sup>, ERA5<sup>43</sup>). For radiative metrics, we use ESA CCI<sup>40</sup> (1982–1999) and ISCCP<sup>38</sup> (1984–1999), as these observations are among the most reliable spanning the majority of the simulated 1979–1999 period. The plots in the panels were created by visualizing the output from ESMValTool (v2.12.0, <https://www.esmvaltool.org/>) using the Psypplot software (v1.5.1, <https://psypplot.github.io/>).

cover over most of Antarctica (by nearly 10%), thereby decreasing the corresponding bias to below 40% cloud cover. As the remaining deviation cannot be entirely explained by the substantial measurement uncertainty in polar regions (up to 15%<sup>47</sup>), this motivates further refinement in future studies.

While the cloud cover root-mean-square error (RMSE) is lower for the automatically tuned (Panel d) compared to the manually tuned (Panel g) ICON-A simulation ( $\Delta\text{RMSE} = -0.98 \text{ Wm}^{-2}$ ), their bias plots are almost indistinguishable as regional differences are typically rather small ( $\leq 10\%$  cloud cover as can be seen in Fig. S7d).

Improvements we find in cloud cover are most valuable if, to some extent, they translate to improved radiative metrics. After all, an accurate representation of LW and SW radiation is crucial for ensuring that the climate model can effectively simulate the Earth's energy budget—a fundamental component for projecting temperature and other climate-relevant parameters. The second column of Fig. 3 reveals that the manually tuned ICON-A (Panel h) generally overestimates outgoing LW radiation, particularly over mountainous regions. This overestimation arises from the model's coarse-resolution grid, which smooths topography and causes mountains appear lower—and therefore warmer—than in reality. The automatically tuned ICON-A (Panel e) shows a reduced outgoing LW radiation globally, except for a localized increase (up to  $15 \text{ Wm}^{-2}$ ) around the Philippines (see also Fig. S7e). ICON-A-MLe further reduces outgoing LW radiation, bringing it closer (with exceptions over the tropics) to observations. Especially over oceans, the increase in cloud cover causes reduced LW radiation now being emitted from cloud tops that are cooler than the ocean's surface. However, ICON-A-MLe tends to underestimate outgoing LW radiation over the tropical western Pacific. This bias stems partly from a (more accurate) increase in cloud ice (see Fig. S8), which forms at higher altitudes and thus emits less LW radiation.

The difference between the SW radiation maps (third column of Fig. 3) from ICON-A-MLe (Panel c) and the automatically tuned ICON-A (Panel f) are highly correlated with those of cloud cover (see also Fig. S7). The largest discrepancies occur in regions with an abundance of marine low-level clouds, where increased cloud coverage leads to more SW radiation being reflected back to space. The increase in cloudiness turns a slight

underestimation of the reflected TOA SW radiation into a slight overestimation and results in a small global RMSE improvement ( $\Delta\text{RMSE} = -0.11 \text{ Wm}^{-2}$ ). Meanwhile, the automatically tuned ICON-A's SW radiation map itself shows improvements ( $\Delta\text{RMSE} = -0.55 \text{ Wm}^{-2}$ ) over the manually tuned version due to its increased marine low-level cloud coverage and consequently increased reflected SW radiation. Across all three metrics—cloud cover, outgoing LW, and reflected SW radiation—ICON-A-MLe consistently achieves the lowest RMSE values, while the manually tuned ICON-A model performs the worst. This underscores the improvement that can be achieved through the combination of the automatic tuning procedure and the data-driven cloud cover equation.

In addition to the three metrics presented in Fig. 3, Table S1 provides a comprehensive comparison of the biases (RMSE and  $R^2$ -values) for other key climate variables across the three simulations. While minor increases in RMSE can occur, the  $R^2$ -values of ICON-A-MLe are generally on par with or better than those of the ICON-A models. Exceptions are the simulation of the LW cloud radiative effect (LWCRE), which is better in the auto-tuned ICON-A model ( $\Delta R^2 = -0.14$ ), and the ice water path, which is slightly better in the manually tuned ICON-A model ( $\Delta R^2 = -0.02$ ). The most striking improvement ( $\Delta R^2 = 0.22$ ) is observed in the liquid water path, defined as the vertically integrated amount of cloud water.

### Isolating the added value of the data-driven cloud cover scheme

While the differences between the baseline ICON-A and the ICON-A-MLe models (Fig. 3), both fully tuned in the same automatic manner, provide indications that the data-driven cloud cover scheme itself contributes to the error reduction, in this section, we quantify more precisely how much of the improvements gains of the previous section we can solely attribute to the data-driven scheme (as opposed to the automatic tuning procedure). To do this, we designed two configurations to test each scheme in an environment optimized for the *other* scheme:

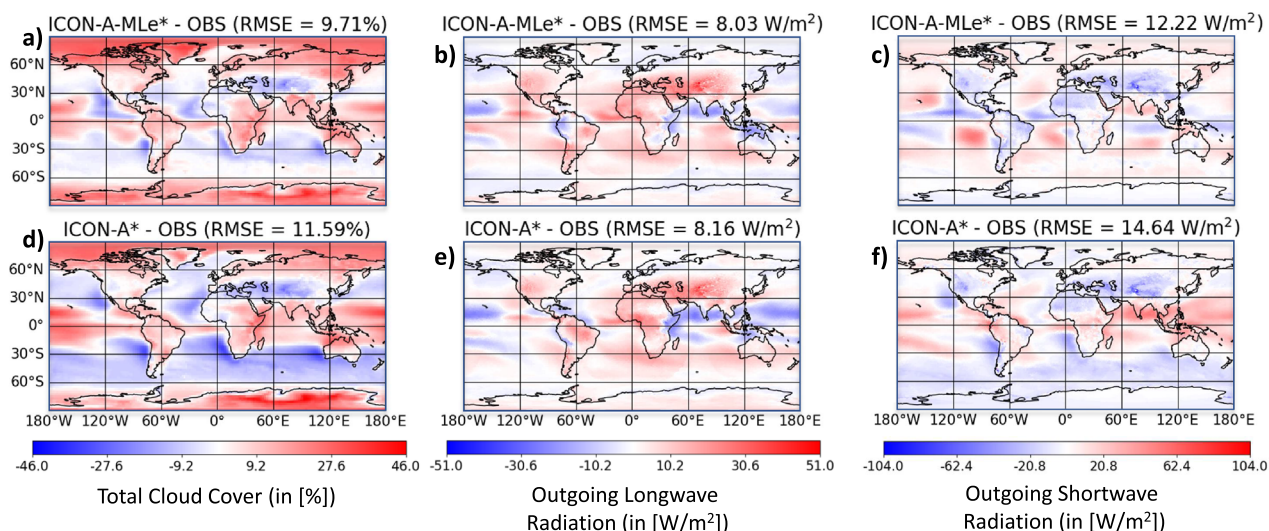
**ICON-A-MLe\*:** We start with the fully tuned ICON-A configuration (the baseline model's optimized environment). We then replace its native cloud scheme with our new data-driven scheme and retune only the parameters of the new scheme. This tests the new scheme in a potentially difficult environment as the tuning procedure is not allowed to mitigate other potential biases present in the climate model.

**ICON-A\*:** Conversely, we start with the fully tuned ICON-A-MLe configuration (the new scheme's optimized environment). We replace the data-driven scheme with the native scheme and retune only the parameters of the native scheme. This tests the native scheme in an environment optimized for the data-driven scheme.

In Fig. 4 we evaluate these two configurations by showing the same statistics as in Fig. 3. The spatial bias patterns remain consistent, confirming the signature behavior of each scheme regardless of the tuning environment. For instance, the ICON-A-MLe models tend to overestimate SW radiation over the Eastern Pacific and the ICON-A models consistently show a strong total cloud cover bias over the Southern Ocean. As opposed to their counterparts, some models seem to trade in more distinct regional biases for reduced global ones (e.g., LW radiation over the Pacific in ICON-A\* or ICON-A-MLe).

Notably, the data-driven scheme performs very well in the ICON-A environment (ICON-A-MLe\*); its global RMSE for total cloud cover (9.71%) and LW radiation (8.03  $\text{Wm}^{-2}$ ) is even slightly lower than in the fully tuned ICON-A-MLe simulation in which all parameters were tuned together in a pursuit to reach a highly dimensional optimum. In contrast, the native scheme struggles significantly in the environment tuned for the data-driven scheme (ICON-A\*), with its SW radiation RMSE increasing to 14.64  $\text{Wm}^{-2}$ .

We can quantify the average impact of the data-driven scheme by averaging its performance gain across both experimental setups. We calculate the average change in RMSE by comparing the new scheme to the old scheme within each of their respective “background” environments:



**Fig. 4.** Evaluation of ICON-A-MLe\* and ICON-A\* 20-year simulations as in Fig. 3.

$$\text{Avg. } \Delta \text{RMSE} = \frac{1}{2} \left( (\text{RMSE}(\text{ICON-A-MLe}^*) - \text{RMSE}(\text{ICON-A})) + (\text{RMSE}(\text{ICON-A-MLe}) - \text{RMSE}(\text{ICON-A}^*)) \right)$$

For the metrics of Fig. 3 this calculation yields

$$\begin{cases} (9.71 - 11.66 + 9.87 - 11.59)/2\% = -1.84\%, & \text{total cloud cover,} \\ (8.03 - 8.62 + 8.34 - 8.16)/2 \text{ Wm}^{-2} = -0.21 \text{ Wm}^{-2}, & \text{longwave radiation,} \\ (12.22 - 11.41 + 11.3 - 14.64)/2 \text{ Wm}^{-2} = -1.27 \text{ Wm}^{-2}, & \text{shortwave radiation.} \end{cases}$$

The negative values for all three metrics confirm that, on average across both tuned environments, the data-driven cloud cover scheme robustly reduces the RMSE compared to the native scheme.

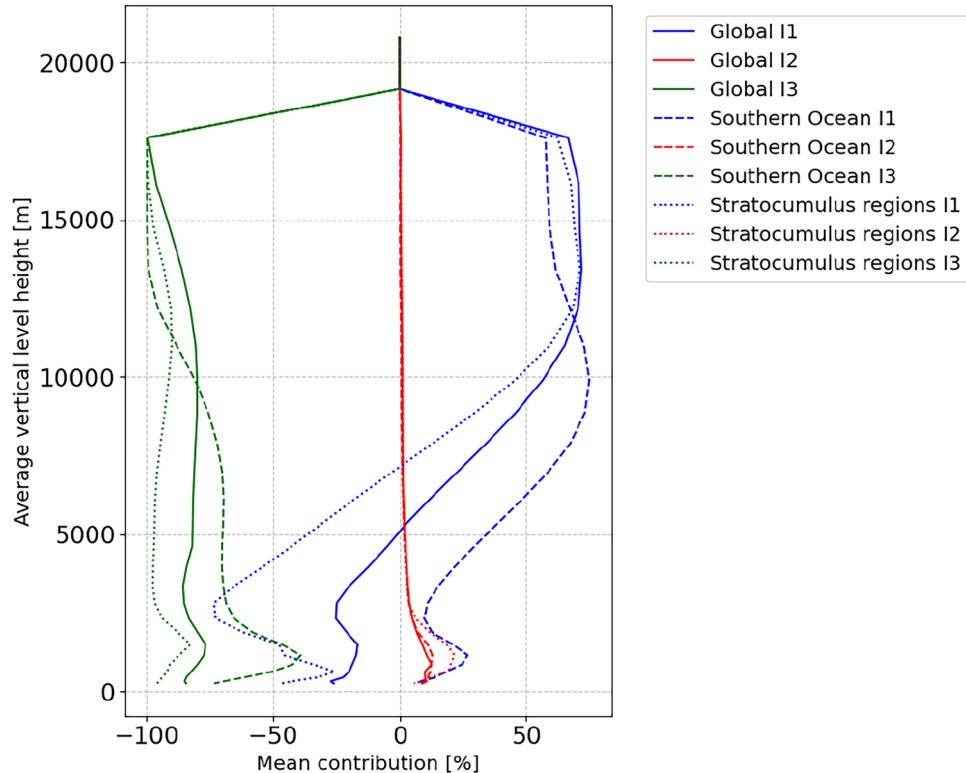
This finding is supported by Table S2, which shows that the data-driven scheme generally improves simulation results across all considered metrics. The most pronounced positive impacts are on liquid water path (LWP), which sees an  $R^2$ -score improvement of  $\approx 0.5$ , and total cloud cover, which benefits from an RMSE reduction of 1.77%. Overall,  $R^2$ -scores improve for five metrics and remain neutral for three. The only metric that degrades is the LWCRE, with an  $R^2$  decrease of 0.09. This appears to be a necessary trade-off, as the baseline model's high  $R^2$ -score for LWCRE was achieved at the cost of underestimating both LWP and total cloud cover.

### Physical drivers of improved cloud cover in ICON-A-MLe

A primary motivation for developing the new analytical parameterization was its physical interpretability<sup>17</sup>. The superior performance of the ICON-A-MLe configuration in maritime regions is not just a product of the optimization methodology but directly linked to its physical formulation (Eq. 3). Figure 5 confirms that the improvements in specific regions correspond directly to the intended physical function of these terms.

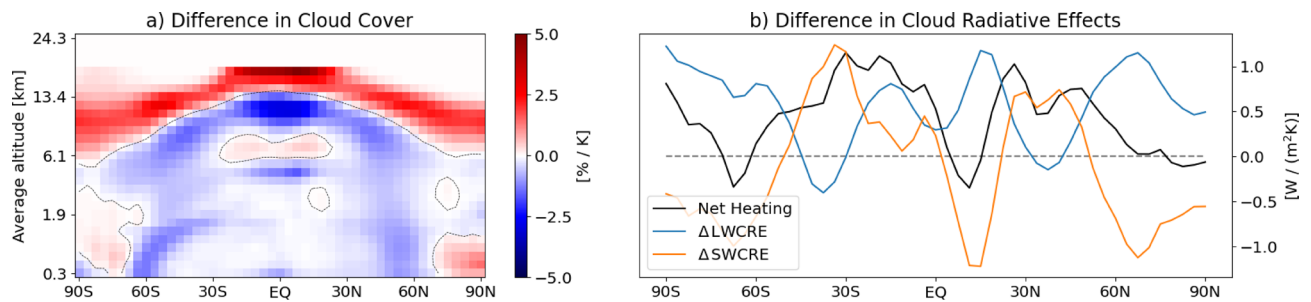
#### Southern Ocean cloudiness (the $I_1$ term)

The most significant reduction in cloud cover biases over the Southern Ocean can be attributed to the  $I_1$  term. As detailed in<sup>17</sup>, this term introduces a more nuanced dependence on temperature ( $T$ ) than the baseline Sundqvist scheme. This is crucial for improving the representation of low-level clouds in this region, which the baseline model struggled to capture. The tuning process optimized the parameters governing  $I_1$ , allowing it to become most active in this specific regime and leading to the documented performance gains.



**Fig. 5.** Vertical profiles of the contributions from each term ( $I_1$ ,  $I_2$ ,  $I_3$ ) of the data-driven cloud cover Eq. (3) to the total cloud fraction on each height level. Values are annual means from an ICON-A-MLe simulation (1979–1980), spatially averaged globally, as well as over two distinct regions. The 'Stratocumulus' regions combine the west coasts of Chile/Peru, Namibia/Angola, California, Morocco, and Australia.





**Fig. 6.** Application of our ICON-A-MLe model in a sea surface temperature increase scenario of +4 K: Panel (a) illustrates the changes in cloud cover per degree of warming as a function of latitude and altitude relative to a control simulation without induced heating. Dashed contour lines are indicating 0 %/K. Panel (b) presents the corresponding changes in longwave (LW) and shortwave (SW) cloud radiative effects (CRE), and the net cloud-induced heating, i.e.,  $\Delta\text{LWCRE} + \Delta\text{SWCRE}$ , as a function of latitude. Both metrics are computed over the last 10 years of a 20-year simulation (1979–1999) and are normalized by the imposed sea surface temperature increase (4 K).

### Subtropical stratocumulus decks (the $I_2$ term)

In contrast, the improved simulation of subtropical stratocumulus decks is driven by the  $I_2$  term. This term is formulated to track the vertical gradient of relative humidity,  $\partial_z \text{RH}$ . It is therefore most sensitive to the strong, low-level temperature inversions that cap these marine boundary layer clouds and are accompanied by a rapid decrease in relative humidity with height. The activation of the  $I_2$  term in these specific regions allows the model to better represent the sharp cloud tops and maintain the stratocumulus layers, correcting known biases in the baseline model.

In summary, the results of the tuning validate the physical hypothesis behind the parameterization's design. The  $I_1$  and  $I_2$  terms function as “dials” that the tuning methodology can adjust, but their design ensures that these adjustments correspond to distinct and physically meaningful processes in specific climate regimes.

### Application of ICON-A-MLe to warmer climates

For a climate model to provide reliable future climate projections, it must respond reasonably to changes in climate conditions. To assess the applicability of ICON-A-MLe to a warmer climate, we increase the SST uniformly by 4 K, following a common proxy for global warming scenarios<sup>8,48,49</sup>. Using the same model setup as in the previous section, we perform a 20-year +4 K simulation with ICON-A-MLe. Figure 6 evaluates only the last 10 years of the simulation to allow for model adjustment to the increased SST.

The +4 K simulation exhibits a notable rise in the cloud top heights across all latitudes, particularly in the tropics (panel a). This lifting is consistent with a deepening of the troposphere, a finding supported by observational evidence<sup>50,51</sup>. The depicted decrease of low-level cloud cover at low to mid-latitudes is a direct consequence of increased SSTs, aligning with many current climate model projections<sup>52</sup>. This effect is more pronounced at 60° S than at 60° N due to the absence of landmasses in the Southern Hemisphere. Over the Arctic, ICON-A-MLe simulates an increase in low-level clouds under the +4 K scenario, in agreement with observations<sup>53</sup> and the CESM1 climate model<sup>54</sup>. This increase is attributed to sea ice loss during fall. Overall, simulated total cloud cover decreases by 1.65% in the +4 K scenario. A reduction in total cloud cover with warming is expected, primarily due to the Clausius–Clapeyron relationship<sup>55</sup>.

The manually tuned ICON-A baseline model also exhibits these characteristics in the +4 K warming scenario (Fig. S9a). While the increase in cloud top altitudes is highly comparable, the decrease of low-level cloudiness at low to mid-latitudes is less pronounced. Furthermore, the increase of low-level Arctic clouds extends to a higher altitude in the troposphere (up to  $\approx 5$  km). The latter contributes to a 1.5% increase in the globally averaged total cloud cover.

Panel (b) presents the difference in CRE between the +4 K and the control simulations, separated into SW and LW components. A positive  $\Delta\text{LWCRE}/\Delta\text{SWCRE}$  indicates that changes in cloudiness warm the troposphere in the +4 K simulation. The panel shows that the LWCRE of clouds generally increases in the +4 K simulation due to higher and thus colder cloud tops, which emit less LW radiation. Additionally, LWCRE changes are sensitive to the altitude at which cloud cover changes occur. The reduction in low-level cloud cover at 60° S/60° N has very little impact on the warming effect of clouds, whereas decreases in mid-level clouds at 40° S, the Equator, and at 40° N lead to a noticeable reduction in  $\Delta\text{LWCRE}$ . Interpreting the SWCRE differences is more nuanced, as CREs are influenced not only by cloud cover but also by cloud optical properties. Given the changes in panel (a), one might expect a symmetric  $\Delta\text{SWCRE}$  pattern around the Equator; however, small differences in cloud cover contribute to a significantly reduced albedo effect and increased heating south of the Equator. The observed decrease in the amplitude of SWCRE at 60° S matches the substantial reduction in marine low-level cloud cover. In the Arctic, changes in SWCRE and LWCRE largely offset each other, while in Antarctica, an increase in high-level clouds enhances the positive net CRE. Overall, cloud cover changes in the +4 K simulation contribute to a net increase in tropospheric heating by  $0.53 \text{ W}/(\text{m}^2 \text{K})$ .

In the manually tuned ICON-A baseline model, the change in LWCRE is nearly identical to that of the ICON-A-MLe model (Fig. S9b). While generally following a similar zonal pattern, the SWCRE between 70° S and 50°

N exhibits a stronger decrease in the +4 K simulation, with a particularly notable reduction of  $\approx 1 \text{ W}/(\text{m}^2 \text{K})$  between  $40^\circ \text{S}$  and  $20^\circ \text{N}$ . This enhanced SW cooling is driven by the baseline model's tendency to retain, and partly even increase, low-level tropical cloud cover in the warmer climate (Fig. S9a). Collectively, these cloud feedbacks contribute to a net global increase in tropospheric cooling of  $0.28 \text{ W}/(\text{m}^2 \text{K})$  in the baseline model.

Both the ICON-A-MLe and the baseline ICON-A models project highly similar, physically plausible changes in global precipitation patterns under the +4K warming scenario (Fig. S10). Specifically, they simulate a strong intensification of rainfall over the central and western Pacific, accompanied by drying over the Maritime Continent and Australia, which resembles El Niño-like conditions. In CMIP3 climate change projections, such conditions are linked to a weakening of the Walker Circulation<sup>56</sup>. The magnitude of the rainfall intensification, averaged over the 10-year period, is up to twice as strong in the ICON-A simulation, peaking at  $2.6 \cdot 10^{-5} \text{ kg}/(\text{m}^2 \text{sK})$  in the Andaman Sea. While the monthly mean precipitation distributions for both models exhibit a similar positive shift of approximately 30%, ICON-A precipitation is generally stronger by roughly  $10^{-4} \text{ kg}/(\text{m}^2 \text{s})$ .

## Conclusion

In this study, we present the first application of a machine-learning based parameterization significantly reducing the persistent cloud cover biases in marine stratocumulus regions (by 24.1–54.2%) and the Southern Ocean (by 75.8%) in a global climate simulation. We show strong online skill of an offline-trained, data-driven yet physically constrained and interpretable parameterization for cloud cover (Eq. 3) alongside a pipeline for automatically calibrating the resulting hybrid climate model with Earth observations. The data-driven cloud cover parameterization, derived via symbolic distillation from high-resolution data, introduces no additional computational burden and maintains the interpretability of the climate model—distinguishing it from many other machine learning-based approaches. The parameterization is implemented into a climate model (ICON-A) and an automatic tuning procedure is used to optimize the performance of the resulting ICON-A-MLe model against observations. Our tuning approach progressively increases the duration of ICON-A-MLe simulations to keep the required computational resources low. The method is characterized by its simplicity, speed (at the expense of quantifying parameter uncertainty), and extensibility to other parameters and metrics.

The tuned ICON-A-MLe model substantially reduces long-standing biases in cloud cover and associated radiative fluxes, particularly in regions dominated by low-level clouds (e.g., the Southern Ocean). More broadly, ICON-A-MLe matches or outperforms the original ICON-A model across a wide range of climate metrics. We use a “scheme-swap” experiment to demonstrate that a significant portion of this improvement is directly attributable to the new cloud scheme itself, independent of the tuning procedure. Furthermore, we leverage the scheme's interpretability to identify physical drivers for cloudiness in the Southern Ocean (relative humidity and temperature) and stratocumulus regions (vertical gradient of relative humidity). When subjected to +4 K surface warming, the model captures physically plausible changes in cloud cover, radiative effects, and precipitation—a crucial step towards improving climate projections.

While for benchmarking purposes this study focuses on ICON 2.6.4, we anticipate that both the tuning pipeline and the data-driven cloud cover parameterization would yield similar improvements in newer ICON versions or other climate models. Additional enhancements to the tuning procedure could include verifying model performance in different seasons by varying initialization dates before extending to longer simulation periods. Additional metrics and parameters—such as those related to climate system dynamics (e.g., orographic drag or surface wind stress;<sup>21</sup>)—could be added to further calibrate the model. Regarding the representation of cloud cover in ICON-A, data-driven equations for cloud inhomogeneity and overlap may be beneficial.

We emphasize that our approach—discovering a new data-driven equation that can be used as a parameterization and tuning the hybrid climate model—is not specific to the parameterization of cloud cover. Given an appropriate training set, this pipeline can be applied to any parameterization of “fast processes” that one aims to model with a low-dimensional analytical equation. Thus, this work presents a framework for improving climate models using data-driven methods in a practical, computationally efficient, and interpretable way towards hybrid climate models with reduced errors that can provide more robust climate projections for mitigation and adaptation assessments.

## Methods

### Tuning the ICON climate model

In this work we use ICON 2.6.4, a recent version of the ICON climate model with parameterizations still closely based on ECHAM physics (as in <sup>21</sup>). We conduct atmosphere-only simulations at a horizontal resolution of 80 km with a model time step of 10 minutes, a radiation time step of 2 h, a vertical grid extending up to 83 km across 47 levels, initialization on Jan 1, 2016, relabeled to 1979 (as in <sup>21</sup>), and boundary data (e.g., aerosol and ozone concentrations) matching the simulated date. As a first step, we establish a reliable baseline that we can compare to and build upon.

#### *Manually tuned baseline*

Our default ICON-A 2.6.4 setup operates on a horizontal grid with increased resolution compared to its original description in <sup>21</sup>, and has seen numerous model changes without being recalibrated. As a consequence, a reasonable top-of-the-atmosphere (TOA) radiative balance (difference between incoming and outgoing radiative fluxes) can only be achieved at the cost of overestimating both TOA net incoming shortwave (SW) and outgoing longwave (LW) radiative fluxes (see Fig. S1). Moreover, total cloud cover is underestimated (global average of 59.3%), which is too small to be deemed acceptable according to <sup>57</sup>. To address these issues, we also manually tuned ICON-A parameters as an additional reference. Specifically, we varied 13 tuning



parameters and utilized approximately 300 simulations from<sup>24</sup>, most of them simulating 10 years, to identify which parameters effectively serve as tuning parameters for the radiative metrics and cloud cover (Fig. S2). Key parameters identified through this process are the coefficient of sedimentation velocity of cloud ice (*cvtfall*), entrainment rate for mid-level convection (*entrmid*), and critical relative humidity for condensation in the upper troposphere (*crt*). By decreasing *cvtfall* from 2.5 to 2.25, allowing ice particles to remain suspended in clouds for longer, slightly increasing *entrmid* from 2e-4 to 2.1e-4 and decreasing *crt* from 0.8 to 0.79, enhancing the formation of high-level clouds at lower relative humidities, we could increase global mean total cloud cover to above 60% (Fig. S1). Additionally, these adjustments yielded SW and LW radiative fluxes at the TOA and a radiative balance, both in close agreement with observations. A comprehensive evaluation of the simulation also demonstrates very reasonable zonal means across various metrics (Fig. S3). By making the convenient yet improper assumption that the impacts of any parameter on a climate metric are linear and independent, we could thus use traditional sensitivity tests to establish a baseline. However, this methodology is both expensive and luck-dependent, highlighting the need for an alternative tuning approach.

#### Automatic tuning pipeline

We hereby introduce a new automatic tuning pipeline (Fig. 1), which we use to tune both the ICON-A baseline model and our ICON-A-MLe model. First, we replace the original cloud cover scheme<sup>58</sup> in ICON-A 2.6.4 by the data-driven equation (Eq. 3) for cloud cover from<sup>17</sup>, which showed superior skill in a comprehensive offline evaluation. We then define which model parameters are to be adjusted in the tuning process. Their initial values match those obtained from the manually tuned baseline. Specifically, we target 24 different tuning parameters (listed in Table S3) with a focus on clouds and convection. Ten of the parameters are part of the cloud cover equation and the other 14 are included in the convection (3), cloud optical properties (8), microphysics (2), and vertical diffusion (1) schemes. Our goal is to tune these parameters such that key climate metrics from model simulations are in good agreement with observations. This can be formulated as an optimization problem. Let  $p \in \mathbb{R}^{24}$  be the vector with the 24 tuning parameters. Our objective is to find an optimal parameter vector  $p_d^*$  that minimizes an objective function,  $J_d$ :

$$p_d^* = \arg \min_p J_d(p).$$

The objective function  $J_d$  (formalized as Eq. (2) in the next subsection) quantifies the deviation of a climate model simulation that is based on a specific realization of these parameters from observational reference data. The optimal parameter vector depends on the choice of climate model (implicitly) and on the duration  $d$  of the model simulation. As optimization method we use the Nelder-Mead (NM) method (also called ‘downhill simplex method’)<sup>35</sup>, which is efficient and does not require differentiation of the objective function. To find a minimum, the NM method moves a simplex through the parameter space, iteratively running the ICON model and evaluating the objective function (i.e., the deviation of a model simulation from the reference data) using the parameter values corresponding to each vertex of the simplex in every step.

Our final goal is to find a parameter set that minimizes  $J_d$ , where  $d$  is on the order of decades. However, to minimize the use of computational resources, we start by simulating only two days. After the NM algorithm has finished tuning the parameters such that the corresponding simulation is as close to the reference data as possible (in practice, when either the objective value is zero or a runtime of 8 h is reached), we extend the simulation duration to a week, then a month, and finally a year. In each iteration with a given ICON simulation duration, the parameters are initialized using the best setting of the previous iteration. The procedure can be formalized as follows:

1. First iteration ( $d_1 = 2$  days): Find  $p_{d_1}^*$  by minimizing  $J_{d_1}(p)$ , starting from an initial guess  $p_0$ .
2. Iterate ( $d_i > d_{i-1}$ ): For subsequent durations ( $d_2 = 1$  week,  $d_3 = 1$  month,  $d_4 = 1$  year), find the new optimum  $p_{d_i}^*$  by minimizing  $J_{d_i}(p)$ , initializing from the result from the previous stage,  $p_{d_{i-1}}^*$ .
3. Minimize  $J_{d_4}(p)$  again, now using  $P$  from Eq. (1) as the initial guess.

In the last iteration, we avoid multiple full NM cycles using expensive year-long simulations. Instead, we first perform a single NM cycle using a limited number (14–20) of year-long simulations, constrained by the 8-hour runtime limit. This yields a set of candidate parameter vectors, from which we select the one that minimizes the objective function, i.e.,  $p_{d_4}^*$ , along with its associated TOA energy balance loss,  $T_{d_4}$ . In parallel, we retain the best parameter set  $p_{d_3}^*$  from the previous NM tuning iteration, along with its corresponding TOA loss in year-long simulations  $T_{d_3}$ . We then construct a new candidate parameter vector  $P$  by extrapolating between  $p_{d_3}^*$  and  $p_{d_4}^*$ . Specifically, we set

$$P = p_{d_3}^* - \frac{T_{d_3}}{T_{d_4} - T_{d_3}}(p_{d_4}^* - p_{d_3}^*), \quad (1)$$

such that the expected TOA loss is minimized under a linear approximation.

A second NM cycle is then performed around this extrapolated parameter vector, again using 14–20 year-long simulations. If any of these runs produce a TOA imbalance in the desirable range of 0.5 to 1 W m<sup>-2</sup>, we select the best among them (with the lowest objective value) even if this value is up to 25% worse than the global minimum. Thereby, we prioritize the TOA balance when sufficiently good solutions are found. Otherwise, we select the parameter set with the lowest objective value. This strategy typically yields high-quality solutions while limiting the number of costly year-long simulations to fewer than 40.

**How do we measure the deviation of a model simulation from given reference data?** First, we define a set of metrics (denoted as  $M$  in Fig. 1) consisting of globally and temporally averaged (excluding the first day) outgoing TOA LW and SW radiation, the balance between incoming and outgoing radiation, SW and LW cloud radiative effects (CRE), total cloud cover, precipitation, ice and liquid water paths, and water vapor path. Additionally, we consider total cloud cover in the Southeastern Pacific [15–45° S, 75–105° W] (averaging at 0.71), a proxy for representing subtropical stratocumulus clouds, that are typically underestimated in climate models. The LWCRE/SWCRE is defined here as the difference between the TOA outgoing LW/SW radiation in clear-sky conditions and the all-sky (actual) outgoing LW/SW radiation. We then assess whether the simulated metrics fall within the respective ranges provided by multiple observational products (e.g., from CERES<sup>37</sup>, GPCP-SG<sup>59</sup>, and ISCCP<sup>38</sup>). The contribution of each metric to the objective value is quantified as the distance between its computed value and the observational range, normalized by the range width to ensure a relative comparison.

Based on the observational products, the expected ranges for outgoing LW and SW radiation are [231, 242] and [100, 110] W/m<sup>2</sup>, respectively, while the net radiation balance should fall within [0.5, 0.9] W/m<sup>2</sup>. Precipitation is expected to range from [2.7, 3.0] mm/day, whereas bounds for the remaining metrics are taken from<sup>60</sup>, Table 5. These bounds are derived from annual means over at least 20 years within the period 1980–2020, and are systematically relaxed (by adding a variability score  $\text{varsc}_{d,m}^{\max}$  to the upper bound  $y_{\text{obs},m}^{\max}$  and subtracting  $\text{varsc}_{d,m}^{\min}$  from the lower bound  $y_{\text{obs},m}^{\min}$ ) for shorter simulations to account for natural variability. Using these components, we define the objective value of a model simulation  $S_d(p)$  of duration  $d$  based on parameter values  $p$  as

$$J_d(p) = \sum_{m \in M} \frac{\max\{S_{d,m}(p) - (y_{\text{obs},m}^{\max} + \text{varsc}_{d,m}^{\max}), 0\} + \max\{y_{\text{obs},m}^{\min} - \text{varsc}_{d,m}^{\min} - S_{d,m}(p), 0\}}{(y_{\text{obs},m}^{\max} + \text{varsc}_{d,m}^{\max}) - (y_{\text{obs},m}^{\min} - \text{varsc}_{d,m}^{\min})}, \quad (2)$$

summing over all metrics  $m \in M$ . This variability scores are estimated by first averaging reference standard ICON-A model outputs over periods matching the current simulation length  $d$ , and by then taking the differences between a metric's most extreme values and its average value

$$\begin{aligned} \text{varsc}_{d,i}^{\max} &= \max\{\overline{S_{\text{ref},i}}^d\} - \overline{S_{\text{ref},i}}, \\ \text{varsc}_{d,i}^{\min} &= \overline{S_{\text{ref},i}} - \min\{\overline{S_{\text{ref},i}}^d\}. \end{aligned}$$

Without such an adjustment of the bounds, shorter simulations would be excessively constrained, limiting the effectiveness of the tuning method. The degree of relaxation varies by metric and simulation length. For example, for global precipitation, the relaxed bounds are [2.34, 3.42], [2.48, 3.23], [2.57, 3.18], and [2.67, 3.03] mm/day for daily, weekly, monthly, and yearly averages, respectively. Note that the required relaxation is relatively small for this metric, as the daily average of global precipitation is already a good approximation of its long-term mean.

**How does the tuning pipeline improve the climate model and affect parameter values?** At the start of the first tuning iteration, using day-long ICON-A-MLe simulations, cloud cover and cloud water are strongly overestimated, negatively affecting the SW radiative budget and inducing errors in the related tuning metrics (Fig. 2). These issues are largely resolved halfway through the iteration, most notably through a substantial reduction in the cloud cover offset (by 32.6%;  $a_1$  in Eq. (3) and Table S3). Towards the end of the iteration,  $a_1$  is increased even above its initial value and a remaining underestimation of the water vapor path is corrected through fine-tuning of all parameter values, decreasing the objective value to zero (see also Fig. S4). The second column of Fig. 2 shows that a 20-year simulation using the parameter values refined in the first iteration already shows substantial improvements in the three radiative metrics, although the radiative balance has not yet been achieved. Such a discrepancy is expected at this stage, as the parameters have only been tuned to produce accurate single-day results so far.

As the automated tuning process progresses, gradually increasing the ICON-A-MLe simulation duration, the model's skill continues to improve. Eventually, the TOA radiative balance—whose observational estimate lies between 0 and 1 W/m<sup>2</sup>—falls within the range of CMIP6 model results. While an objective value of zero is quickly attained in tuning iterations using week- and month-long simulations, in year-long simulations it is decreased from 2.42 to 1.77 (see also Fig. S5). A final loss of zero is difficult to attain as it requires a full reconciliation of high total cloud cover over the Southeast Pacific with an accurate TOA radiative budget.

Table S3 details the parameter modifications due to tuning, with the parameters of the data-driven cloud cover Eq. (3) listed first. Besides the reduction in the cloud cover offset, notable changes include a slight increase in cloud cover sensitivity to relative humidity and temperature, as well as a 42% decrease in the numerical stabilizer  $\epsilon$ , which discourages the formation of unrealistic condensate-free clouds. Dependencies on cloud water/ice and vertical relative humidity gradients remain largely unchanged, reinforcing their role in representing thin marine stratocumulus clouds in the model. Adjustments in other parameterizations include setting the Prandtl number slightly above 1 (1.02), modifying cloud droplet number concentrations at low altitudes, and increasing cloud inhomogeneity factors for most cloud types.

In general, we find that it is beneficial to modify many parameters slightly, which is difficult to achieve manually. We find that those parameters that undergo significant changes already do so in the first tuning iteration with day-long simulations (e.g.,  $\text{cn2sea}$  is changed from 80 to 41.71). In subsequent iterations, parameter adjustments typically become smaller and tend to stabilize. Exceptions include the numerical stabilizer, which is further reduced in the month-long iteration (from 0.80 to 0.62).

The overall computational cost of tuning the ICON-A-MLe model automatically amounts to approximately 400 node-hours (using 12 AMD EPYC 7763 MILAN CPU nodes in parallel), running 741/29/32/28 ICON-A-MLe simulations in the tuning iterations using daily/weekly/monthly/yearly simulations, respectively.

**Automatic tuning of the ICON-A baseline** To ensure maximum comparability between the ICON-A-MLe and ICON-A models, we apply the same automatic tuning approach to the ICON-A baseline. For comparison, tuning the ICON-A baseline does not yield any single dominant parameter modification. Some parameters that were decreased in the ICON-A-MLe tuning (e.g., mid-level entrainment rate, autoconversion of cloud droplets to rain) are instead increased (see Table S4), suggesting a dependency on the cloud cover scheme and demonstrating Nelder-Mead's ability to adjust parameters accordingly. A notable modification is an increase of the critical relative humidity at the top (by 3.4%) in the iteration with weekly simulations, which reduces the simulated amount of high clouds, correcting an issue of overestimated outgoing SW radiation (reducing it by  $1.8 \text{ W/m}^2$ ). In the case of the ICON-A baseline we found that by nudging the critical relative humidities for cloud formation (from 0.99 to 0.968 at the top and from 0.59 to 0.665 at the bottom of the atmosphere) before the second NM cycle, we end up with a slightly better ICON-A model configuration (TOA loss of zero and an objective value of 3.52 instead of 3.78 without the adjustment).

Throughout the tuning process, the challenge of balancing high total cloud cover over the Southeast Pacific with an accurate TOA radiative budget remains more evident in the traditional atmospheric model. Therefore, the total computational cost is higher, amounting to approximately 600 node-hours, with 432/327/158/36 ICON-A simulations in the four tuning iterations.

### Data-driven cloud cover equation

The data-driven equation for cloud cover from<sup>17</sup>, discovered with symbolic regression<sup>18</sup> on coarse-grained data from storm-resolving DYAMOND simulations<sup>61,62</sup>, can be formulated as follows:

$$f(\text{RH}, T, \partial_z \text{RH}, q_c, q_i) = I_1(\text{RH}, T) + I_2(\partial_z \text{RH}) + I_3(q_c, q_i), \quad (3)$$

a function of relative humidity (RH), its vertical derivative ( $\partial_z \text{RH}$ ), temperature (T), cloud water ( $q_c$ ), cloud ice content ( $q_i$ ) with

$$\begin{aligned} I_1(\text{RH}, T) &= a_1 + a_2(\text{RH} - \overline{\text{RH}}) + a_3(T - \overline{T}) \\ &\quad + \frac{a_4}{2}(\text{RH} - \overline{\text{RH}})^2 + \frac{a_5}{2}(T - \overline{T})^2(\text{RH} - \overline{\text{RH}}) \\ I_2(\partial_z \text{RH}) &= a_6^3 \left( \partial_z \text{RH} + \frac{3a_7}{2} \right) (\partial_z \text{RH})^2 \\ I_3(q_c, q_i) &= \frac{-1}{q_c/a_8 + q_i/a_9 + \epsilon}. \end{aligned}$$

After automatically tuning Eq. (3) online within ICON-A, its attained parameter values are

$$\{a_1, \dots, a_9, \epsilon\} = \{0.118, 1.234, -0.0265 \text{ K}^{-1}, 5.65, 1.56 \cdot 10^{-3} \text{ K}^{-2}, \\ 591.68 \text{ m}, 2.22 \text{ km}^{-1}, 1.47 \text{ mg/kg}, 0.344 \text{ mg/kg}, 0.615\}$$

(see also Table S3). The values for the reference relative humidity  $\overline{\text{RH}}$  (0.6025) and temperature  $\overline{T}$  (257.06 K) are fixed.

### Implementation of the data-driven scheme in ICON-A

The data-driven cloud cover parameterization is implemented as a diagnostic, three-dimensional cloud fraction scheme within the ICON-A model (version 2.6.4)<sup>21</sup>. It is computed at every model time step for each atmospheric grid cell, using the local thermodynamic state (e.g., T, RH, and their vertical gradients) from that cell as inputs for Eq. (3). This scheme replaces the model's native Sundqvist-based diagnostic cloud fraction calculation<sup>58</sup>. It interacts with other model components in the following ways:

1. Cloud microphysics: The model's existing one-moment microphysics scheme continues to prognose the liquid and ice water mixing ratios ( $q_l, q_i$ )<sup>63</sup>. Our new scheme provides the cloud fraction ( $C$ ) required to compute the tendencies of liquid and ice water and water vapor in each grid cell.
2. Radiation: These 3D fields ( $C$  and grid-mean condensates) are passed to the PSrad radiation scheme<sup>64</sup>. The radiation scheme's internal vertical overlap logic (a standard maximum-random assumption) is unchanged. PSrad operates on these 3D inputs to compute the final 2D radiative fluxes at the top of the atmosphere and at the surface.

### Data availability

Algorithms, code, and data for reproducing our results are available at our GitHub page [https://github.com/EyringMLClimateGroup/grundner25\\_iconaml\\_automatic\\_tuning/tree/main](https://github.com/EyringMLClimateGroup/grundner25_iconaml_automatic_tuning/tree/main) and preserved at <https://doi.org/10.5281/zenodo.15194060>.

Received: 18 August 2025; Accepted: 14 November 2025

Published online: 13 December 2025

## References

1. Stensrud, D. J. *Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models* (Cambridge University Press, 2009).
2. Eyring, V. et al. Pushing the frontiers in climate modelling and analysis with machine learning. *Nat. Clim. Change* **14**(9), 916–928 (2024).
3. Eyring, V., Gentine, P., Camps-Valls, G., Lawrence, D. M. & Reichstein, M. AI-empowered next-generation multiscale climate modelling for mitigation and adaptation. *Nat. Geosci.* **17**, 963–971. <https://doi.org/10.1038/s41561-024-01527-w> (2024).
4. Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. Could machine learning break the convection parameterization deadlock? *Geophys. Res. Lett.* **45**(1), 5742–5751 (2018).
5. Grundner, A., Beucler, T., Gentine, P., Iglesias-Suarez, F., Giorgetta, M. A., & Eyring, V. Deep learning based cloud cover parameterization for ICON. *J. Adv. Model. Earth Syst.* **14**(12), e2021MS002959 (2022).
6. Heuer, H., Schwabe, M., Gentine, P., Giorgetta, M. A., & Eyring, V. Interpretable multiscale machine learning-based parameterizations of convection for ICON. *J. Adv. Model. Earth Syst.* **16**(8), e2024MS004398 (2024).
7. O’Gorman, P. A. & Dwyer, J. G. Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *J. Adv. Model. Earth Syst.* **10**(10), 2548–2563 (2018).
8. Rasp, S., Pritchard, M. S., & Gentine, P. Deep learning to represent subgrid processes in climate models. *Proc. Natl. Acad. Sci.* **115**(39), 9684–9689. <https://doi.org/10.1073/pnas.1810286115> (2018).
9. Reichstein, M. et al. Deep learning and process understanding for data-driven Earth system science. *Nature*. **566**(7743), 195–204 (2019).
10. Grundner, A. Data-driven cloud cover parameterizations for the ICON earth system model using deep learning and symbolic regression. Accessed: 2024-10-16. PhD thesis. Universität Bremen, Jan. 2024. 10.26092/elib/2821. <https://media.suub.uni-bremen.de/handle/elib/7739>.
11. Bracco, A., Brajard, J., Dijkstra, H. A., Hassanzadeh, P., Lessig, C., & Monteleoni, C. Machine learning for the physics of climate. *Nat. Rev. Phys.* **7**(1), 6–20 (2025).
12. Frezat, H., Le Sommer, J., Fablet, R., Balarac, G., & Lguensat, R. A posteriori learning for quasi-geostrophic turbulence parametrization. *J. Adv. Model. Earth Syst.* **14**(11), e2022MS003124 (2022).
13. Christopoulos, C. et al. Online learning of entrainment closures in a hybrid machine learning parameterization. *J. Adv. Model. Earth Syst.* **16**(11), e2024MS004485 2024MS004485, e2024MS004485. <https://doi.org/10.1029/2024MS004485> (2024).
14. Chen, G., Wang, W.-C., Yang, S., Wang, Y., Zhang, F., & Wu, K. A neural network-based scale-adaptive cloud-fraction scheme for GCMs. *J. Adv. Model. Earth Syst.* **15**(6), e2022MS003415 (2023).
15. Morcrette, C. J. et al. Scale-aware parameterization of cloud fraction and condensate for a global atmospheric model machine-learned from coarse-grained kilometer-scale simulations. *Authorea Preprints* (2024).
16. S. Rasp. The optimization dichotomy: Why is it so hard to improve climate models with machine learning? (2021). <https://raspste.phan.github.io/blog/optimization-dichotomy/> (accessed 16 Oct 2024).
17. Grundner, A., Beucler, T., Gentine, P., & Eyring, V. Data-driven equation discovery of a cloud cover parameterization. *J. Adv. Model. Earth Syst.* **16**(3), e2023MS003763 (2024).
18. M. Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl. In: arXiv preprint [arXiv:2305.01582](https://arxiv.org/abs/2305.01582) (2023).
19. Huntingford, C., Nicoll, A. J., Klein, C., & Ahmad, J. A. Potential for equation discovery with AI in the climate sciences. *Earth Syst. Dyn.* **16**(2), 475–495 (2025).
20. Lai, C.-Y., Hassanzadeh, P., Sheshadri, A., Sonnewald, M., Ferrari, R., & Balaji, V. Machine learning for climate physics and simulations. *Annu. Rev. Condensed Matter Phys.* **16** (2024).
21. Giorgetta, M. A. et al. ICON-A, the atmosphere component of the ICON earth system model: I. Model description. *J. Adv. Model. Earth Syst.* **10**(7), 1638–1662. <https://doi.org/10.1029/2017ms001233> (2018).
22. Neelin, J. D., Bracco, A., Luo, H., McWilliams, J. C., & Meyerson, J. E. Considerations for parameter optimization and sensitivity in climate models. *Proc. Natl. Acad. Sci.* **107**(50), 21349–21354 (2010).
23. Xie, S., Ma, H.-Y., Boyle, J. S., Klein, S. A., & Zhang, Y. On the correspondence between short-and long-time-scale systematic errors in CAM4/CAM5 for the year of tropical convection. *J. Clim.* **25**(22), 7937–7955 (2012).
24. Bonnet, P. et al. Tuning a climate model with machine-learning based emulators and history matching. *EGU sphere* **2024**, 1–32 (2024).
25. Roach, L. A., Tett, S. F., Mineter, M. J., Yamazaki, K. & Rae, C. D. Automated parameter tuning applied to sea ice in a global climate model. *Clim. Dyn.* **50**, 51–65 (2018).
26. Eyring, V. et al. Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.* **9**, 1937–1958 (2016).
27. Righi, M. et al. Earth system model evaluation tool (ESMValTool) v2.0—technical overview. *Geosci. Model Dev.* **13**(3), 1179–1199. <https://doi.org/10.5194/gmd-13-1179-2020> (2020).
28. 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**(7), 3383–3438. <https://doi.org/10.5194/gmd-13-3383-2020> (2020).
29. Lauer, A. et al. Earth system model evaluation tool (ESMValTool) v2.0—diagnostics for emergent constraints and future projections from Earth system models in CMIP. *Geosci. Model Dev.* **13**(9), 4205–4228. <https://doi.org/10.5194/gmd-13-4205-2020> (2020).
30. Weigel, K. et al. Earth system model evaluation tool (ESMValTool) v2.0—diagnostics for extreme events, regional and impact evaluation, and analysis of Earth system models in CMIP. *Geosci. Model Dev.* **14**(6), 3159–3184. <https://doi.org/10.5194/gmd-14-3159-2021> (2021).
31. Andela, B. et al. ESMValCore. Version v2.11.0. July 2024. <https://doi.org/10.5281/zenodo.3387139>.
32. Andela, B. et al. ESMValTool. Version v2.11.0. July 2024. <https://doi.org/10.5281/zenodo.3401363>.
33. Lauer, A., Bock, L., Hassler, B., Jöckel, P., Ruhe, L., & Schlund, M. Monitoring and benchmarking Earth system model simulations with ESMValTool v2. 12.0. *Geosci. Model Dev.* **18**(4), 1169–1188 (2025).
34. Schlund, M. et al. Evaluation of native Earth system model output with ESMValTool v2.6.0. *Geosci. Model Dev.* **16**(1), 315–333. <https://doi.org/10.5194/gmd-16-315-2023> (2023).
35. Nelder, J. A. & Mead, R. A simplex method for function minimization. *Comput. J.* **7**(4), 308–313. <https://doi.org/10.1093/comjnl/7.4.308> (1965).
36. Gelaro, R. et al. The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *J. Clim.* **30**(14), 5419–5454 (2017).
37. Loeb, N. G. et al. Clouds and the earth’s radiant energy system (CERES) energy balanced and filled (EBAF) top-of-atmosphere (TOA) edition-4.0 data product. *J. Clim.* **31**(2), 895–918 (2018).
38. Zhang, Y. & Rossow, W. B. Global radiative flux profile data set: Revised and extended. *J. Geophys. Res. Atmos.* **128**(5), e2022JD037340 (2023).
39. Karlsson, K.-G. et al. CLARA-A2.1: CM SAF cLoud, Albedo and surface RAdiation dataset from AVHRR data—Edition 2.1. 2020. 10.5676/EUM\_SAF\_CM/CLARA\_AVHRR/V002\_01. [https://wui.cmsaf.eu/safira/action/viewDoiDetails?acronym=CLARA\\_AVHRR\\_V002\\_01](https://wui.cmsaf.eu/safira/action/viewDoiDetails?acronym=CLARA_AVHRR_V002_01).



40. Stengel, M. et al. Cloud property datasets retrieved from AVHRR, MODIS, AATSR and MERIS in the framework of the Cloud\_cci project. *Earth Syst. Sci. Data* **9**(2), 881–904 (2017).
41. Platnick, S. et al. The MODIS cloud products: Algorithms and examples from Terra. *IEEE Trans. Geosci. Remote Sens.* **41**(2), 459–473 (2003).
42. Heidinger, A. K., Foster, M. J., Walther, A., & Zhao, X. The pathfinder atmospheres—extended AVHRR climate dataset. *Bull. Am. Meteorol. Soc.* **95**(6), 909–922 (2014).
43. Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**(730), 1999–2049 (2020).
44. Sommer, P. S. *psypilot*: Interactive data visualization with Python. <https://doi.org/10.5281/zenodo.593798>. <https://codebase.helmholtz.cloud/psypilot/psypilot>.
45. Zelinka, M. D. et al. Causes of higher climate sensitivity in CMIP6 models. *Geophys. Res. Lett.* **47**(1), e2019GL085782 (2020).
46. Muhlbaier, A., McCoy, I. L., & Wood, R. Climatology of stratocumulus cloud morphologies: microphysical properties and radiative effects. *Atmos. Chem. Phys.* **14**(13), 6695–6716 (2014).
47. Liu, X., He, T., Sun, L., Xiao, X., Liang, S., & Li, S. Analysis of daytime cloud fraction spatiotemporal variation over the arctic from 2000 to 2019 from multiple satellite products. *J. Clim.* **35**(23), 7595–7623 (2022).
48. Clark, S. K. et al. Correcting a 200 km resolution climate model in multiple climates by machine learning from 25 km resolution simulations. *J. Adv. Model. Earth Syst.* **14**(9), e2022MS003219 (2022).
49. Kochkov, D. et al. Neural general circulation models for weather and climate. *Nature* **632**(8027), 1060–1066 (2024).
50. Marvel, K. et al. External influences on modeled and observed cloud trends. *J. Clim.* **28**(12), 4820–4840. <https://doi.org/10.1175/JCLI-D-14-00734.1> (2015).
51. Norris, J. R., Allen, R. J., Evan, A. T., Zelinka, M. D., O'Dell, C. W., & Klein, S. A. Evidence for climate change in the satellite cloud record. *Nature* **536**(7614), 72–75. <https://doi.org/10.1038/nature18273> (2016).
52. Ceppi, P., Briant, F., Zelinka, M. D., & Hartmann, D. L. Cloud feedback mechanisms and their representation in global climate models. *Wiley Interdiscip. Rev. Clim. Change* **8**(4), e465 (2017).
53. Kay, J. E. et al. Recent advances in Arctic cloud and climate research. *Curr. Clim. Change Rep.* **2**, 159–169. <https://doi.org/10.1007/s40641-016-0051-9> (2016).
54. Morrison, A. L., Kay, J. E., Frey, W. R., Chepfer, H., & Guzman, R. Cloud response to Arctic sea ice loss and implications for future feedback in the CESM1 climate model. *J. Geophys. Res. Atmos.* **124**(2), 1003–1020. <https://doi.org/10.1029/2018JD029142> (2019).
55. Mendoza, V., Pazos, M., Garduño, R., & Mendoza, B. Thermodynamics of climate change between cloud cover, atmospheric temperature and humidity. *Sci. Rep.* **11**(1), 21244. <https://doi.org/10.1038/s41598-021-00474-1> (2021).
56. Vecchi, G. A. & Soden, B. J. Global warming and the weakening of the tropical circulation. *J. Clim.* **20**(17), 4316–4340 (2007).
57. Mauritsen, T. et al. Tuning the climate of a global model. *J. Adv. Model. Earth Syst.* **4**(3) (2012).
58. Sundqvist, H., Berge, E., & Kristjánsson, J. E. Condensation and cloud parameterization studies with a mesoscale numerical weather prediction model. *Monthly Weather Rev.* (1989).
59. Adler, R. F. et al. The Global Precipitation Climatology Project (GPCP) monthly analysis (new version 2.3) and a review of 2017 global precipitation. In: *Atmosphere* **9.4** (2018), p. 138. <https://doi.org/10.3390/atmos9040138>. <https://www.mdpi.com/2073-4433/9/4/138>.
60. Lauer, A., Bock, L., Hassler, B., Schröder, M., & Stengel, M. Cloud climatologies from global climate models—A comparison of CMIP5 and CMIP6 models with satellite data. *J. Clim.* **36**(2), 281–311. <https://doi.org/10.1175/JCLI-D-22-0181.1> (2023). <https://journals.ametsoc.org/view/journals/clim/36/2/JCLI-D-22-0181.1.xml>.
61. Stevens, B. et al. DYAMOND: the Dynamics of the Atmospheric general circulation Modeled On Non-hydrostatic Domains. *Progr. Earth Planet. Sci.* **6**(1), 1–17 (2019).
62. Stephan, C. C. et al. Atmospheric energy spectra in global kilometre-scale models. *Tellus A Dyn. Meteorol. Oceanogr.* **74**(1) (2022).
63. Lohmann, U. & Roeckner, E. Design and performance of a new cloud microphysics scheme developed for the ECHAM general circulation model. *Clim. Dyn.* <https://doi.org/10.1007/BF00207939> (1996).
64. Pincus, R. & Stevens, B. Paths to accuracy for radiation parameterizations in atmospheric models. *J. Adv. Model. Earth Syst.* **5**(2), 225–233. <https://doi.org/10.1002/jame.20027> (2013).

## Acknowledgements

This work used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID bd1179.

## Author contributions

A.G., T.B. and V.E. designed the research; A.G. and J.S. performed the research and analyzed the data; A.L. and M.S. created the concept and ESMValTool code used for the evaluation of the simulations; A.G. wrote the paper with contributions from all co-authors.

## Funding

Open Access funding enabled and organized by Projekt DEAL. A.G., J.S., M.S., and V.E. received funding for this study from the European Research Council (ERC) Synergy Grant “Understanding and modeling the Earth System with Machine Learning (USMILE)” under the Horizon 2020 research and innovation programme (Grant agreement No. 855187) and from the Horizon Europe project “Artificial Intelligence for enhanced representation of processes and extremes in Earth System Models (AI4PEX)” (Grant agreement ID: 101137682). T.B. received support from AIPEX, funded by the Swiss State Secretariat for Education, Research and Innovation (SERI, Grant No. 23.00546). V.E. was additionally supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the Gottfried Wilhelm Leibniz Prize awarded to V.E. (reference no. EY 22/2-1). Open Access funding enabled and organized by Projekt DEAL.

## Declarations

## 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/s41598-025-29155-3>.

**Correspondence** and requests for materials should be addressed to A.G.

**Reprints and permissions information** is available at [www.nature.com/reprints](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) 2025