https://doi.org/10.5194/egusphere-2024-2508
Preprint. Discussion started: 9 August 2024

Tuning a Climate Model with Machine-learning based Emulators and History Matching

Pauline Bonnet[1,*], Lorenzo Pastori[1,*], Mierk Schwabe[1], Marco A. Giorgetta[2], Fernando Iglesias-Suarez[1], and Veronika Eyring[1,3]

[1]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany
[2]Max Planck Institute for Meteorology Hamburg (Germany)
[3]University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany
[*]These authors contributed equally to this work.

Correspondence: Pauline Bonnet (pauline.bonnet@dlr.de) and Lorenzo Pastori (lorenzo.pastori@dlr.de)

Abstract. In climate model development, *tuning* refers to the important process of adjusting uncertain free parameters of subgrid-scale parameterizations to best match a set of Earth observations such as global radiation balance or global cloud cover. This is traditionally a computationally expensive step as it requires a large number of climate model simulations to create a Perturbed Parameter Ensemble (PPE), which is increasingly challenging with increasing spatial resolution and complexity of climate models. In addition, this manual tuning relies strongly on expert knowledge and is thus not independently reproducible. Here, we develop a Machine Learning (ML)-based tuning method with the goal to reduce subjectivity and computational demands. This method consists of three steps: (1) creating a PPE of limited size with randomly selected parameters, (2) fitting an ML-based emulator to the PPE and generate a large PPE with the emulator, and (3) shrinking the parameter space with history matching. We apply this method to the Icosahedral Nonhydrostatic Weather and Climate Model (ICON) for the atmosphere to tune for global radiative and cloud properties. With one iteration of this method, we achieve a model configuration yielding a global top-of-atmosphere net radiation budget in the range of $[0,1]$ W/m^2, and global radiation metrics and water vapor path consistent with the reference observations. Furthermore, the resulting ML-based emulator allows to identify the parameters that most impact the outputs that we target with tuning. The parameters that we identified as mostly influential for the physics output metrics are the critical relative humidity in the upper troposphere and the coefficient conversion from cloud water to rain, influencing the radiation metrics and global cloud cover, together with the coefficient of sedimentation velocity of cloud ice, having a strong non-linear influence on all the physics metrics. The existence of non-linear effects further motivate the use of ML-based approaches for parameter tuning in climate models.

1 Introduction

Climate and Earth system models are developed and continuously improved to understand the behaviour of the Earth system and to project climate change (Tebaldi et al., 2021). Due to their complexity as well as constraints on computational resources, the resolution of climate models is relatively coarse, so that a number of key processes occur on scales smaller than the model grid scale. These non-resolved processes, such as convection, radiation, turbulence, cloud microphysics, and gravity waves, are described statistically for each grid cell through so-called parameterizations, which are a cause of biases and uncertainties in climate projections (Gentine et al., 2021) due to uncertainties in their formulation and in the selection of the underlying free parameters. To constrain the values of the free parameters involved in the parameterizations, tuning is an important step in the development of climate models (Hourdin et al., 2017), where these parameters are adjusted such that the outputs of the climate model reproduces the observed states of the Earth system reasonably well.

Model tuning is typically a very time-consuming and computationally expensive step. It has to be conducted for all components of a climate model (such as atmosphere, ocean and land) and for the coupled model (see for instance the tuning of the coupled ICON Earth System Model by Jungclaus et al. (2022)).

Traditionally, tuning in climate models is done manually, i.e., the parameters are changed individually (or few at a time) in a sequential manner, with expert knowledge guiding the successive choices in the tuning of the parameters (Hourdin et al., 2017; Mauritsen et al., 2012; Schmidt et al., 2017; Giorgetta et al., 2018; Mignot et al., 2021). Such manual approaches may retain some form of subjectivity, and are therefore hard to replicate. There is also the risk of neglecting interactions among the processes affected by the changed parameters, which may lead to compensating errors, e.g., a model's low climate sensitivity might be paired with weak aerosol cooling, resulting in an apparent match with historical data but potentially inaccurate future projections.

In this work we investigate how machine learning (ML) techniques can help addressing the aforementioned challenges faced in model tuning, using the atmospheric component of the ICON model (Giorgetta et al., 2018) as an example. In recent years, ML-based *automatic* tuning methods have been widely investigated. These methods intend to tune the climate models in fewer manual steps for the user compared to fully manual approaches, and aim to improve the accuracy and reproducibility of parameter tuning by formulating it as an optimization problem amenable to numerical treatment. A number of mathematical tools have been developed to tackle inverse problems such as model tuning. The one we focus on in this work belongs to the family of Bayesian approaches, which have been the most commonly used ones in climate model tuning. We note that this is not the only possible choice, and refer to (Zhang et al., 2015) for more details on other possible choices. Bayesian approaches have been implemented to tune models of different complexity, including the Lorenz '63 and '96 models (Cleary et al., 2021), a single column model (Couvreux et al., 2021), cloud related parameters in the LMDZ 6A atmospheric model (Hourdin et al., 2021), an idealized global circulation model (Dunbar et al., 2021), the ECHAM6.3- HAM2.3 atmospheric model for a small number of parameters by Watson-Parris et al. (2021), an intermediate complexity climate model (Mansfield and Sheshadri, 2022), the CLM5 land model (Dagon et al., 2020), the Community Atmosphere Model (CESM2-CAM6) experimenting tuning of 45 parameters targeting three radiation metrics and the liquid water path (Eidhammer et al., 2024), for constraining parameter values in the coupled non-flux climate model (HadCM3) (Williamson et al., 2013), or for uncertainty quantification (Williamson et al., 2017; Hourdin et al., 2023).

Automatic parameter tuning methods formulate the problem as the minimization of a suitably defined distance between the model outputs and observation-based reference datasets. In a Bayesian setting, this is solved by an iterative and efficient exploration of the space of the parameters being tuned, which is enabled by the construction of an ML-based surrogate or emulator of the climate model that aims at approximating the climate model outputs at much cheaper computational costs.

https://doi.org/10.5194/egusphere-2024-2508
Preprint. Discussion started: 9 August 2024

This typically consists in iterating the following steps: (1) generate a perturbed parameter ensemble (PPE), i.e., an ensemble of climate model simulations obtained by sampling configurations of tuning parameters within the valid parameter ranges, (2) train a computationally cheap ML-based emulator on the PPE output to approximate the parameter-to-output relationship, and (3) use the emulator for a denser sampling of the parameter space, and shrink the space of allowed parameter configurations to the most promising one, i.e., the parameters most likely yielding a tuned version of the climate model. A commonly adopted method for selecting promising parameter configurations is history matching (Williamson et al., 2013, 2017). History matching aims at minimizing the number of required model simulations in the search of optimal parameters, by balancing the need of sampling from unexplored regions in the parameter space and the need of exploiting the information gained from the already sampled configurations (e.g., sampling close to already promising parameter configurations). These three steps are repeated until the optimal parameter values, or distributions thereof (Watson-Parris et al., 2021), have been found.

Building on previous climate model tuning efforts (Giorgetta et al., 2018; Couvreux et al., 2021), here we develop an ML-based tuning approach for the atmospheric component of the Icosahedral Nonhydrostatic Weather and Climate Model (ICON-A) (ICON, 2015; Zängl et al., 2014). We tune ICON-A in a two-step approach, first focusing on global radiative and cloud properties, referred to as *physics* outputs (Giorgetta et al., 2018), and then on outputs related to atmospheric circulation properties, referred to as *dynamics* outputs (Giorgetta et al., 2018). We apply our ML-based approach to the first tuning step, constructing emulators that accurately capture the relationship between the changed parameters and the physics outputs, and showing that history matching converges towards observational references in a few iterations. The ML-based tuning of the physics outputs serves as the basis for the second tuning step targeting the dynamics outputs. For this step we follow the approach of Giorgetta et al. (2018) by generating a PPE and selecting the best performing model configurations, where our criterion for evaluating the model's performance puts the highest priority on achieving a nearly balanced global annual net radiation flux at top of the atmosphere (TOA). Our results are compared to the manually tuned version of the ICON-A model that was presented in Giorgetta et al. (2018); Crueger et al. (2018). In the remainder of the paper, we refer to this tuned ICON version as *ICON-aes-1.3*.

The article is organized as follows. We first introduce the ICON-A model, the ML-based tuning method and the reference datasets used in this study in Section 2. We then present the results of the ML-based tuning approach for ICON-A in Section 3 and conclude in Section 4.

2 Methods

2.1 ICON-A modelling framework

The Icosahedral Nonhydrostatic Weather and Climate Model (ICON) is a modelling framework for Climate and Numerical Weather prediction developed jointly by the German Weather Service (DWD) and the Max Planck Institute for Meteorology (MPI-M) (ICON, 2015; Zängl et al., 2014). We use ICON's atmospheric component (ICON-A) (Zängl et al., 2014; Giorgetta et al., 2018), version 2.6.4, and conduct AMIP experiments with the icosahedral grid *R2B5* (≈ 80 km in the horizontal, for details see Table 1 in Giorgetta et al. (2018)) with an implicitly coupled land model. The top height of the atmospheric model is 83 km with 47 full vertical levels and numerical damping starting at 50 km. Subgrid-scale processes are described by parameterizations and include radiative effects, moist convection, vertical diffusion, cloud microphysics, cloud cover, and orographic and non-orographic gravity waves (Giorgetta et al., 2018). The time steps used in the model simulations are one hour for the radiation scheme and six minutes for the atmospheric scheme. For our PPEs we run ICON-A for one year spin up (1979) and then for one year for tuning physics outputs (1980). We then run the model for one year spin up (1979) and then for ten years (1980-1989) for the dynamics outputs, as described in the following sections.

2.2 Parameters and Outputs

The first step to ML-based tuning, as for manual tuning, is to select the tuning parameters and output metrics that are to be fitted. Our choice of the metrics is informed by the manual tuning of the ICON model by Giorgetta et al. (2018) and Crueger et al. (2018). There, the authors worked on model versions preceding ICON-aes-1.3.00, which resulted from their work, with a coarser resolution R2B4 of ≈ 160 km, 47 vertical layers resolving the atmosphere up to a height of 83 km, and time steps of two hours for the radiation scheme and ten minutes for the atmospheric scheme.

Table 1 reports the output metrics that we focus on in this study, which represent global radiative and cloud properties and are referred to as the *physics* outputs. These physics output metrics are all global and multi-year averages. In particular, as shown in Table 1, we use the annual average over 1980 in our PPEs (apart from our last PPE, as discussed later), and compare it with the multi-year averages of the reference datasets reported in Table 1.

The output metrics related to atmospheric circulation properties, the *dynamics* outputs, are given in Table 2. There, the zonal mean velocity at 60° North and South at 10 hPa serves as proxy for the representation of high latitude jets. The surface downward, eastward wind stress mean over the North Atlantic Ocean and the Southern Ocean (defined in the AR6 database (Iturbide et al., 2020)) are proxies for the forcing on the ocean surface. These dynamics output metrics are multi-year averages. In particular, as shown in Table 2, we use the average over the period 1980-1989 in our PPEs, and compare it to the multi-year averages of the reference datasets reported in Table 2.

Following Giorgetta et al. (2018), the parameterizations we select for tuning for the physics outputs are moist convection, vertical diffusion, cloud microphysics and cloud cover. In Table 3 we report the parameters from these parameterizations (which we therefore refer to as physics parameters) which we select for our tuning experiment. The parameterizations we select for tuning for the dynamics outputs are the orographic and non-orographic gravity waves schemes. In Table 4 we report the parameters from these parameterizations (referred to as dynamics parameters) which we select for our tuning experiment.

2.3 Reference datasets

To tune ICON-A we use reference values for the output metrics from Earth observations and reanalysis data. As in Giorgetta et al. (2018), the main goal here is *to obtain a slightly positive global annual mean downward net radiation flux at the top of the atmosphere (TOA), between 0 and 1 W/m^2, based on a net shortwave flux and an outgoing longwave radiation close to observational estimates*. For the two radiation fields (rsdt-rsut) and rlut (see Tab. 1 for definitions), the typical interval [240 W/m^2, 241 W/m^2] is used as a reference value, as estimated in (Giorgetta et al., 2018), following observational datasets

Physics outputs metrics	Spatial average	Averaging period	Ref. datasets	Target range
TOA net shortwave (SW) radiation (rsdt-rsut)	Global (references and PPEs)	1980	Giorgetta et al. (2018)	[240, 241] W/m^2
TOA net longwave (LW) radiation (rlut)	Global (references and PPEs)	1980	Giorgetta et al. (2018)	[-241, -240] W/m^2
TOA radiation balance (rsdt-rsut-rlut)	Global (references and PPEs)	1980	Giorgetta et al. (2018)	[0, 1] W/m^2
Cloud cover (clt)	Global (references and PPEs)	1982-1991 1980-1989 (1980 for PPEs)	CLARA-AVHRR V002 ESACCI-Cloud AVHRR-AMPM-fv3.0	62.7 % 65.1 %
Water vapor path (prw)	Global (references and PPEs)	1980-1989 (1980 for PPEs)	ERA5	[24.1] kg/m^2

Table 1. Physics outputs together with respective observational datasets (CERES-EBAF (NASA/LARC/SD/ASDC, 2019) and ERA5 (Dee et al., 2011)) and target ranges used in this work. All the outputs in this table are globally averaged (for both the reference datasets and the ICON-A simulations we conduct). The averaging period used for both reference datasets and our simulations (PPEs) is reported in the third column. TOA stands for "top of the atmosphere".

Dynamics Output metrics	Spatial average	Averaging period	Ref. datasets	Target range
Zonal wind velocity (ua)	60° North at 10 hPa (references and PPEs)	1980-1989 (references and PPEs)	ERA5, MERRA2, ERA-Interim	(10.94, 11.15, 10.94) m/s
Zonal wind velocity (ua)	60° South at 10 hPa (references and PPEs)	1980-1989 (references and PPEs)	ERA5, MERRA2, ERA-Interim	(32.77, 34.03, 33.15) m/s
Surface downward eastward wind stress (tauu)	North-Atlantic Ocean (NAO) (references and PPEs)	1980-1989 (references and PPEs)	ERA5, MERRA2, ERA-Interim	(2.947e-3, 5.395e-3, 3.645e-3) N/m^2
Surface downward eastward wind stress (tauu)	Southern Ocean (SOO) (references and PPEs)	1980-1989 (references and PPEs)	ERA5, MERRA2, ERA-Interim	(0.1367, 0.1413, 0.1359) N/m^2

Table 2. Dynamics outputs together with respective observational datasets (ERA5 (Hersbach et al., 2020)) used in this work. The North Atlantic Ocean (NAO) region and the Southern Ocean (SOO) region are those defined in the AR6 database (Iturbide et al., 2020).

(CERES-EBAF-Ed4.0, 2000-2016) and Kato et al. (2013); Loeb et al. (2009). For cloud cover, we use CLARA-AVHRR

https://doi.org/10.5194/egusphere-2024-2508
Preprint. Discussion started: 9 August 2024

Physics parameters with corresponding ranges			Parameterization
Average entrainment rate for midlevel convection	entrmid	[2e-5, 3e-4]	Moist convection
Average entrainment rate for penetrative convection	entrpen	[2e-5, 6e-4]	Moist convection
Average entrainment rate for cumulus downdrafts	entrdd	[5e-5, 6e-4]	Moist convection
Characteristic adjustment time scale [s]	cmftau	[2e2, 1e4]	Moist convection
Neutral limit Prandtl number	pr0	[5e-1, 1.2]	Vertical diffusion
Critical relative humidity parameter at the upper troposphere	crt	[5e-1, 9e-1]	Cloud cover
Fractional convective mass flux across the top of cloud	cmfctop	[1e-2 , 2e-1]	Moist convection
Coefficient for determining conversion from cloud water to rain	cprcon	[1.5e-5, 3.5e-4]	Moist convection
Coefficient of autoconversion of cloud ice to snow	ccsaut	[0.2 , 4]	Cloud microphysics
Minimum in-cloud water mass mixing ratio in mixed phase clouds	csecfrl	[1.0e-5 , 1.0e-4]	Cloud microphysics
Coefficient of sedimentation velocity of cloud ice	cvtfall	[0.2 , 4]	Cloud microphysics
Critical relative humidity at surface	crs	[7.26e-1, 9.9e-1]	Cloud cover
Lower limit of scaling factor for saturation mixing ratio in layer below inversion	csatsc	[0.35, 1,05]	Cloud cover

Table 3. Tuning parameters related to physics parameterizations alongside the corresponding name in the ICON source code (second column from left), the range of values tested (third column from left), and the corresponding parameterization scheme they belong to (right column). The range of the parameters was inferred from the default value of the parameters given in the source code of ICON-A version 2.6.4.

Dynamics parameters with associated ranges			Parameterization
Coefficient for orographic gravity wave drag	gkdrag	[0.002, 0.28]	Sub-grid scale orographic effects
Coefficient for low level blocking	gkwake	[0.001, 0.09]	Sub-grid scale orographic effects
Root mean square gravity wave wind at the emission level	rmscon	[0.647, 1.079]	Atmospheric gravity wave effects
Minimum difference "SSO peak height - SSO mean height" [m]	gpicmea	[20,60]	Sub-grid scale orographic effects
Minimum standard deviation of SSO height [m]	gstd	[5,15]	Sub-grid scale orographic effects

Table 4. Tuning parameters related to dynamics parameterizations alongside the corresponding name in the ICON source code (second column from left), the range of values tested (third column from left), and the corresponding parameterization scheme they belong to (right column). SSO stands for "subgrid-scale orography".

(Karlsson et al., 2020) and ESACCI-CLOUD (Stengel et al., 2017), and for the water vapour path, we use ERA5 (Hersbach et al., 2020) (see Section A for time series of these observational datasets). For the dynamics outputs, we use ERA5, ERA-Interim (Dee et al., 2011) and MERRA2 (Gelaro et al., 2017). We refer the reader to Appendix A for the time series of some of the observational products used in this work.

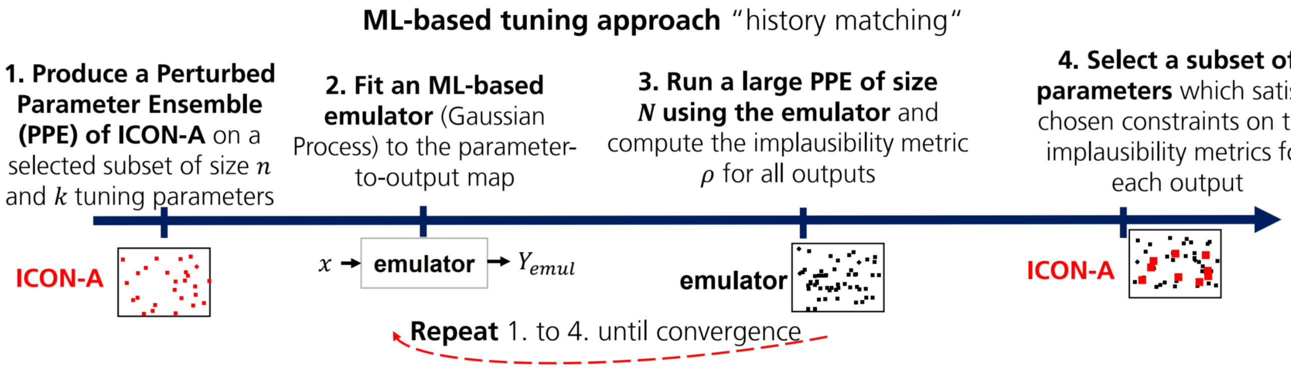

ML-based tuning approach "history matching"
1. Produce a Perturbed Parameter Ensemble (PPE) of ICON-A on a selected subset of size n and k tuning parameters
2. Fit an ML-based emulator (Gaussian Process) to the parameter-to-output map
3. Run a large PPE of size N using the emulator and compute the implausibility metric ρ for all outputs
4. Select a subset of parameters which satisfy chosen constraints on the implausibility metrics for each output
ICON-A
x
emulator
Y_emul
emulator
ICON-A
Repeat 1. to 4. until convergence

Figure 1. Schematic of the history matching technique used for the ML-based parameter tuning of ICON-A.

2.4 ML-based tuning approach

Our ML-based tuning method is built on the history matching technique (Williamson et al., 2013, 2017), and follows a similar workflow as in (Couvreux et al., 2021). The goal is to minimize the distance between the model and the observational data for the outputs introduced before, by exploring the parameter space. In performing this exploration, history matching aims at finding a balance between exhaustively exploring, or sampling, the parameter space, and minimizing the number of samples required for it. Since in our case each sample corresponds to an expensive climate model simulation, we consider this method particularly well suited for our tuning task. We now outline the steps of that approach (see also Fig. 1):

1. Draw an initial Latin Hypercube (LHC) sampling of size n for k tuning parameters. Using LHC sampling, all parameters are simultaneously changed and the different samples fill the k-dimensional parameter space (within the allowed ranges specified in Tables 3 and 4) approximately uniformly. Typically, n is chosen as $n \approx 10\,k$ (Loeppky et al., 2009). Using these selected parameters, generate a PPE of ICON-A runs. The PPE consists of n members, or runs, one for each sampled parameter configuration $\boldsymbol{x}_i$ (with $i = 1, ..., n$). For each run, we calculate all the output metrics described before. This results in sets of input-output training pairs $\mathcal{T}_Y = \{\boldsymbol{x}_i, Y_{\text{model}}(\boldsymbol{x}_i)\}_{i=1,...,n}$, one set per output metric Y (e.g., global TOA radiation balance).

2. Fit an emulator to the generated PPE, i.e., to the training sets $\mathcal{T}_Y$ for all the output metrics Y of interest. For a given metric Y, the emulator evaluated on a configuration of tuning parameters $\boldsymbol{x}$ returns $Y_{\text{emul}}(\boldsymbol{x})$, the approximation to the true model output metric $Y_{\text{model}}(\boldsymbol{x})$. Our choice for the model emulator is Gaussian process (GP) regression (Rasmussen and Williams, 2005). GPs are models typically used in Bayesian regression tasks, and are very well suited for our case since (i) they have only few parameters, hence require relatively little training data for fitting, and (ii) they by construction return the uncertainty associated to their prediction, which is measured by the variance $\text{Var}(Y_{\text{emul}}(\boldsymbol{x}))$. This is a central quantity used in the steps below. Further details on the choice of the GP are given in Appendix B. In our implementation, we train one GP per model output.

https://doi.org/10.5194/egusphere-2024-2508
Preprint. Discussion started: 9 August 2024

3. Generate a large PPE of size N (typically ranging from 10^5 to 10^6) using the trained GP emulator. For each emulator run, calculate the implausibility measure ρ for each metric Y, with reference value Y^0 (from observations or re-analysis data) as:

$$\rho(Y_{\text{emul}}(\boldsymbol{x}), Y^0) = \frac{|Y^0 - Y_{\text{emul}}(\boldsymbol{x})|}{\sqrt{\text{Var}(Y_{\text{emul}}(\boldsymbol{x}))}} \ . \tag{1}$$

4. Select n parameter configurations that satisfy the following constraints on the outputs (see Table 1 and Table 2 for outputs definitions):

   - $\rho(Y_{\text{emul}}(\boldsymbol{x}), Y^0) < \rho_1$: for the three physics metrics TOA shortwave radiation, TOA longwave radiation, and TOA net incoming radiation,
   - $\rho(Y_{\text{emul}}(\boldsymbol{x}), Y^0) < \rho_2$: for the two other physics metrics cloud cover and liquid water path, and the five dynamics metrics.

   The choice of a smaller threshold for the three radiation metrics is necessary in order to give a higher weight to the constraint on a balanced TOA radiation than on the other metrics. We use $\rho_2 = 2\rho_1$. The value of ρ_1 is automatically adjusted in order to select only n parameter sets out of the ensemble of size N.

5. Generate a new size n PPE with ICON-A for the parameter ensemble defined in the previous step, and repeat steps 3, 4 and 5 until an optimal parameter configuration is reached.

The criterion to select the optimal parameter configuration is based on a weighted distance of the model output metrics from their reference value, with the highest weight given to the global TOA net radiation balance, that is our main tuning goal.

We implement the GP emulator in Python using scikit-learn (https://scikit-learn.org/stable/), and used the built-in routines to optimize the GP parameters at each iteration of the above procedure (see details in Appendix B). In this work, we measure the performance of the GP regression model via the R^2 value, which for a given output Y is defined as:

$$R^2(Y) = 1 - \frac{\overline{(Y_{\text{emul}} - Y_{\text{model}})^2}}{\text{Var}(Y_{\text{model}})} \ , \tag{2}$$

where $\overline{(Y_{\text{emul}} - Y_{\text{model}})^2}$ denotes the mean squared error of the emulator over a set of testing parameters, and $\text{Var}(Y_{\text{model}})$ the variance of the true model output over the same test set.

We divide the tuning of ICON-A into two steps. In the first step we use the ML approach to tune the parameterizations related to the physics outputs. The model configuration resulting from the application of history matching in this step serves as basis for the second step where physics and dynamics parameters and metrics are tuned simultaneously. Furthermore, the PPEs generated in the first step are used for performing a sensitivity analysis of the physics parameters, using Sobol indices (Saltelli et al., 2007), to compute the influence of each parameter on the physics outputs. This makes it possible to identify and keep only the most influential parameters for the second step. We refer the reader to Appendix C for more details.

The second tuning step incorporates both physics and dynamics outputs simultaneously, following the manual tuning process by Giorgetta et al. (2018). This two-step approach is justified given the faster temporal response and shorter equilibrium timescale of physics outputs compared to the dynamics outputs.

https://doi.org/10.5194/egusphere-2024-2508
Preprint. Discussion started: 9 August 2024

3 Results

3.1 Summary of the generated PPEs

The PPEs generated in this work are summarized in Table 5. PPE_1 to PPE_4 are generated for the tuning of the physics output metrics from single-year ICON-A runs (1980) after a one year spin-up. PPE_1 is generated from an LHC sampling of size 30 on the (physics) parameter set:

$$\mathcal{P}_{\mathrm{p1}} = \{\mathrm{entrpen}, \mathrm{entrmid}, \mathrm{entrdd}, \mathrm{cmftau}, \mathrm{crt}, \mathrm{pr0}\} \ . \tag{3}$$

PPE_2 is produced by applying history matching on the results of PPE_1. For PPE_3 we perform a new LHC sampling on an extended parameter set:

$$\mathcal{P}_{\mathrm{p2}} = \{\mathrm{cmfctop}, \mathrm{cprcon}, \mathrm{ccsaut}, \mathrm{csecfrl}, \mathrm{cvtfall}, \mathrm{crt}, \mathrm{pr0}\} \ , \tag{4}$$

in order to increase the globally averaged cloud cover, which is consistently lower than the observational references in the previous PPEs. The sensitivity analysis shown in Appendix C shows that the new parameters in $\mathcal{P}_{\mathrm{p2}}$ do indeed have a strong influence on global cloud cover. For generating PPE_3 and PPE_4, the values of the parameters in $\mathcal{P}_{\mathrm{p1}}$ that are not present in $\mathcal{P}_{\mathrm{p2}}$ are fixed to their best value from PPE_2 (see the right column of Table 5 and the magenta star in Fig. 2 and Fig. 3). PPE_4 is produced by applying history matching on the results of PPE_3.

In PPE_5 we then address also the tuning of dynamics outputs by varying physics and dynamics parameters simultaneously in the parameter set:

$$\mathcal{P}_{\mathrm{pd}} = \{\mathrm{entrmid}, \mathrm{cvtfall}, \mathrm{crt}, \mathrm{crs}, \mathrm{csatsc}, \mathrm{rmscon}, \mathrm{gkdrag}, \mathrm{gkwake}, \mathrm{gpicmea}, \mathrm{gstd}\} \ , \tag{5}$$

and keeping the other parameters fixed to their best values in PPE_2 (see the right column of Table 5 and the magenta star in Fig. 2 and Fig. 3). PPE_5 consists of ten-year ICON-A simulations from 1980 to 1989 (after a one year spin-up). The selection of the parameters in $\mathcal{P}_{\mathrm{pd}}$ was aided by a sensitivity analysis based on Sobol indices, which we present in Appendix C. We now move on to discuss the results from the PPEs.

3.2 ML-based tuning of physics outputs with history matching

In this section we present the results of the tuning of the physics parameters. We start by considering PPE_1 and PPE_2. As explained before, PPE_2 is generated by applying history matching after having trained a GP emulator on the outputs of PPE_1. The constructed GP emulator in this case has a good predictive performance (measured by an average R^2 score of 0.81, as discussed in more details in Section 3.2.1 below), and can therefore accurately guide the parameter choices for PPE_2. Thanks to this, the application of only one iteration of history matching to PPE_1 is already sufficient to generate configurations in PPE_2 that achieve a balanced TOA radiation. This is demonstrated in panel (a) of Fig. 2, which shows the net short-wave (SW) versus the net long-wave (LW) TOA radiation for PPE_1 and PPE_2. There, we can clearly see that after history matching on PPE_1, PPE_2 can achieve configurations that match or get close to the observational ranges denoted by the green triangle

PPE	Parameters changed	Size	Description	Outputs	Fixed parameters
PPE_1	$\mathcal{P}_{\text{p1}} = \{$entrpen, entrmid, entrdd, cmftau, crt, pr0$\}$	30	LHC sampling of $\mathcal{P}_{\text{p1}}$	physics	cmfctop (0.1), cprcon (2.5e-4), ccsaut (2.0), csecfrl (1.5e-5), cvtfall (2.5), csatsc (0.7), crs (0.968)
PPE_2	$\mathcal{P}_{\text{p1}}$	30	History matching from PPE_1	physics	
PPE_3	$\mathcal{P}_{\text{p2}} = \{$cmfctop, cprcon, ccsaut, csecfrl, cvtfall, crt, pr0$\}$	30	LHC sampling of $\mathcal{P}_{\text{p2}}$	physics	entrpen (9.295e-5), entrmid (2.2504e-4), entrdd (1.766e-4), cmftau (2114.6), csatsc (0.7), crs (0.968)
PPE_4	$\mathcal{P}_{\text{p2}}$	30	History matching from PPE_3	physics	
PPE_5	$\mathcal{P}_{\text{pd}} = \{$entrmid, cvtfall, crt, crs, csatsc, rmscon, gkdrag, gkwake, gpicmea, gstd$\}$	80	LHC sampling of $\mathcal{P}_{\text{pd}}$	physics and dynamics	entrpen (9.295e-5), entrdd (1.766e-4), cmftau (2114.6), pr0 (0.93168), ccsaut (2.0), csecfrl (1.5e-5)

Table 5. Summary of perturbed parameters ensembles (PPEs) generated in this work. The PPEs have been sequentially generated from 1 to 5. PPE_3 is obtained from an LHC sampling of parameter set $\mathcal{P}_{\text{p2}}$, where the parameters in $\mathcal{P}_{\text{p1}}$ not included in $\mathcal{P}_{\text{p2}}$ are kept fixed to their best values from PPE_2 (and listed in the right column), which are then used further in PPE_4 and PPE_5.

(and to ICON-aes-1.3). The convergence of the output metrics towards their reference values can also be observed in panel (b) of Fig. 2, for the other two physics output metrics (global cloud cover versus water vapor path) for PPE_1 and PPE_2. There, the distribution of the PPE_2 outputs is converging towards the observational references (green markers). In Fig. 2, the magenta star marks the selected best performing model configuration in PPE_2. Following Giorgetta et al. (2018), our criterion for evaluating the model performance prioritizes the global radiation metrics, in particular the net TOA radiation budget, over cloud cover and water vapor path. The selected is the only one falling within the observational range for both radiation metrics (green triangle in panel (a)). The convergence of history matching towards the observational references can also be seen in the distribution of the sampled parameters for the two PPEs (Fig. 3).

Fig. 2 panel (b) shows that global cloud cover still remains lower than the observational data (of approximately 1% compared to CLARA-AVHRR, and 3% compared to ESACCI), despite PPE_2 yielding a slightly higher cloud cover (closer to the observed range) than PPE_1.

Therefore, in the next generation of PPEs, we extend the parameter set $\mathcal{P}_{\text{p2}}$ to contain parameters to which cloud cover is more sensitive. Parameter set $\mathcal{P}_{\text{p2}}$ is used to generate PPE_3 with LHC sampling. A GP emulator is then trained on the outputs of PPE_3. The constructed GP emulator in this case also has a good predictive performance (measured by an average R^2 score of 0.75, as discussed in more details in Section 3.2.1 below), and we therefore use it for performing history matching and generating PPE_4. Also in this case history matching is shrinking the space of promising parameter configurations and

https://doi.org/10.5194/egusphere-2024-2508
Preprint. Discussion started: 9 August 2024

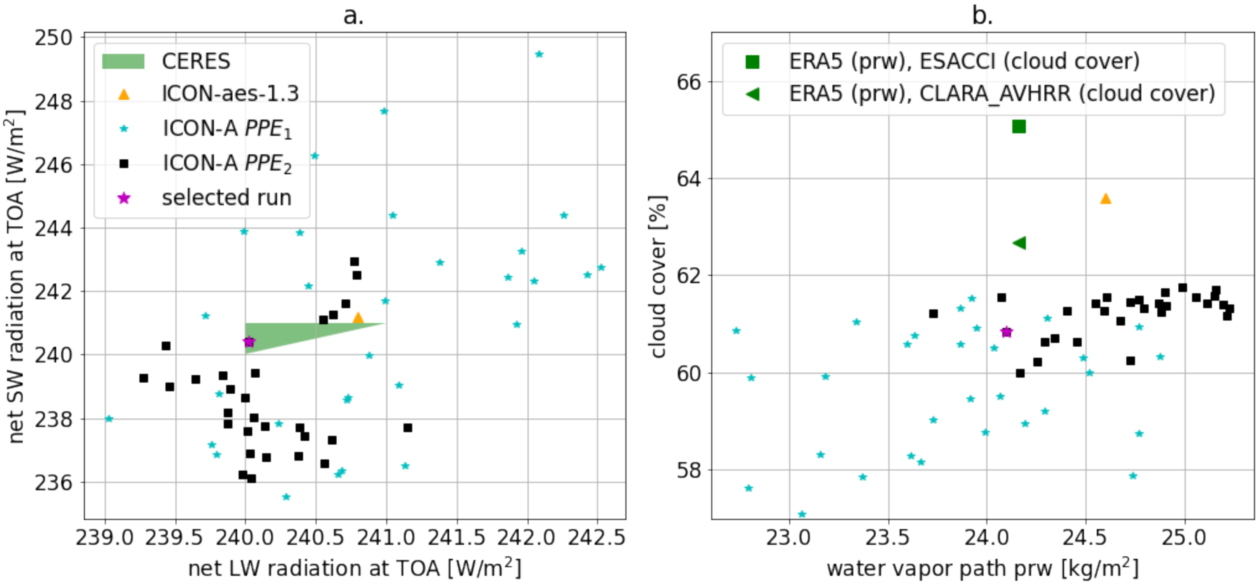


Figure 2. Physics output variables for PPE_1 (blue stars) and PPE_2 (black squares) compared to ICON-aes-1.3 (orange triangle) and observational datasets (green). Signs of convergence of history matching are visible already after one iteration (the distribution of the members of PPE_2 is slightly shifted towards higher cloud cover values and narrower). The magenta star marks the best performing configuration from PPE_2 (see right column of Table 5), used in the generation of the subsequent PPEs.

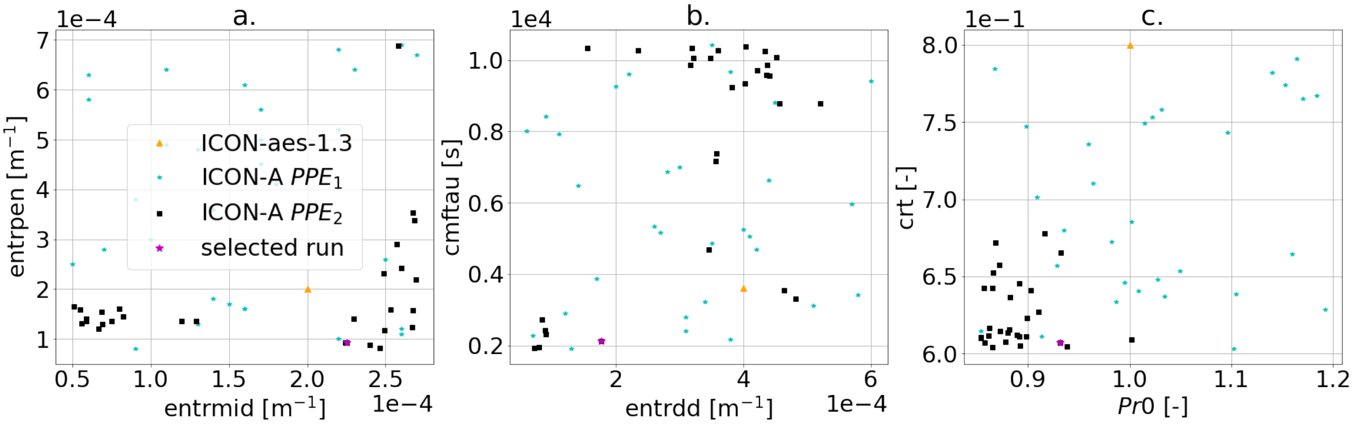


Figure 3. Sampled parameter values for PPE_1 (blue stars) and PPE_2 (black squares) compared to ICON-aes-1.3 (orange triangle). For each panel, two parameters are plotted on the two axes (see Table 3). Signs of convergence of history matching are visible already after one iteration (in the distribution of the members of PPE_2 being slightly shifted and narrower). The magenta star marks the best performing configuration from PPE_2 (see also right column of Table 5 for values), used in the generation of the subsequent PPEs.

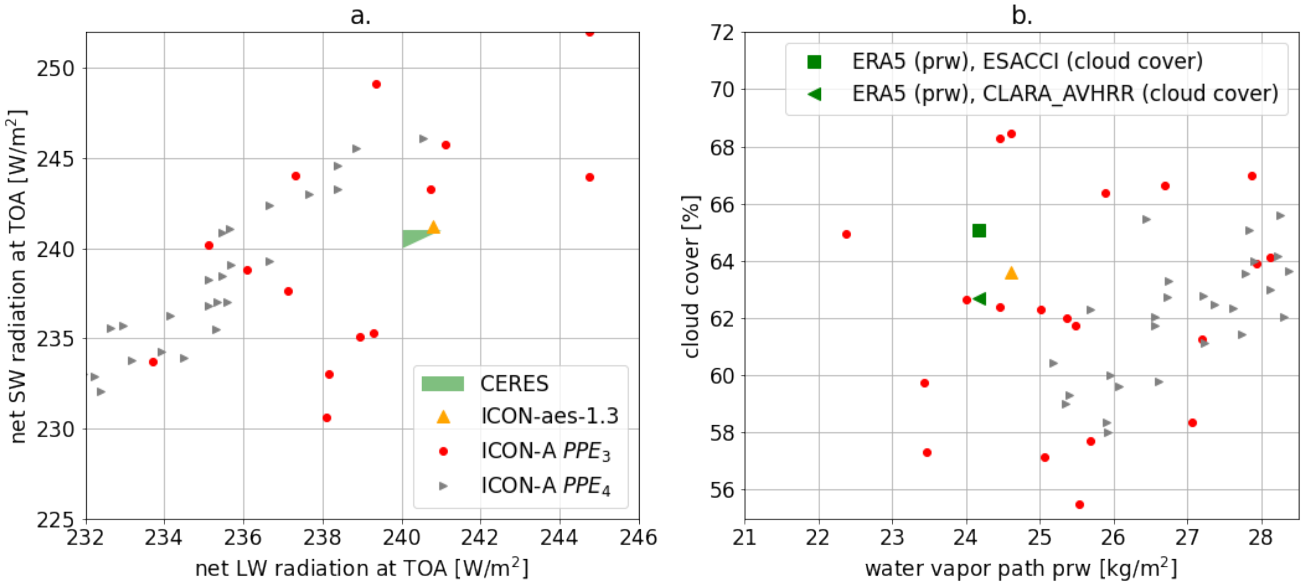


Figure 4. Physics output variables for PPE_3 (red circles) and PPE_4 (grey triangles) compared to ICON-aes-1.3 (orange triangle) and observational datasets. Also here, signs of convergence of the outputs to their observational values can be seen (in the distribution of the members of PPE_4 being slightly shifted and narrower).

the related output distribution. This can be seen in Fig. 4, where we show the distribution of the radiation metrics (in panel (a)), and the of global cloud cover versus water vapor path (in panel (b)) for both PPE_3 and PPE_4 (we refer the reader to Appendix D for plots of the related parameters distributions). While the new parameter set $\mathcal{P}_{\mathrm{p}2}$ allows us to reach a global cloud cover consistent with observations, we also see that the spread of the PPE outputs is almost doubled compared to that of the previous PPEs. This increased spread also potentially increases the number of history matching iterations to converge towards the observational references. Given the high computational costs of generating these PPEs, we therefore use the best performing model configuration sampled so far, which belongs to PPE_2.

3.2.1 Performance of the GP emulator

We now analyze the performance of the GP emulator for the physics outputs considered. We refer the reader to Appendix B for details on Gaussian processes and the choice of the underlying hyperparameters. In Table 6, we show the average performance (R^2 score) of the GP emulators trained on the PPEs used for the tuning of the physics parameters (corresponding to PPE_1, PPE_2, PPE_3, PPE_4). The value reported in Table 6 is the average R^2 over all the five physics output metrics (defined in Table 1), and is computed using a 5-fold cross validation (https://scikit-learn.org/stable/). From these values, we conclude that the constructed emulators are indeed able to approximate the ICON-A physics outputs, which is also reflected in the fact that

https://doi.org/10.5194/egusphere-2024-2508
Preprint. Discussion started: 9 August 2024

PPE used for training	GP-emulator R^2-score
PPE_1	0.82
PPE_1 + PPE_2	0.79
PPE_3	0.75
PPE_3 + PPE_4	0.81

Table 6. Performance of the GP-emulator on PPE_1 to PPE_4. The R^2 value reported here is the average R^2 of the emulators for all physics variables (see Table 1). For each emulator, the R^2 is calculated via 5-fold cross validation on the training set (PPE points).

history matching shows signs of convergence already after the first iteration, as shown in the previous section. The number of PPE samples required for the GP regression to achieve the reported R^2 score is shown in Fig. 5.

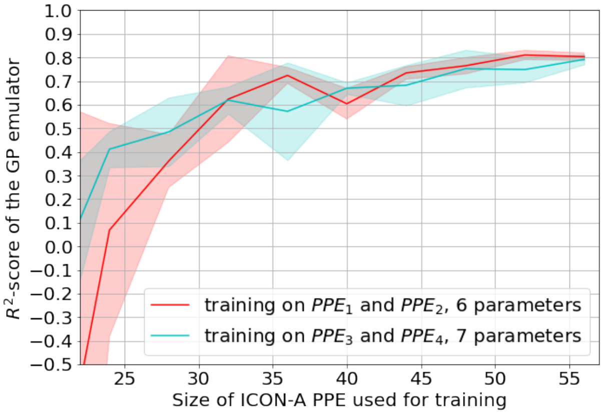


Figure 5. Average R^2 score of the physics outputs emulators, as a function of the number N of PPE samples used for training. For each N tested, five randomly drawn samples of size N were drawn from the entire set of ICON PPEs of size 60. The R^2 score is calculated for each size-N sample, and the mean (solid lines) and standard deviation (shaded areas) are estimated from these scores on the five samples. The red curve shows the R^2 for emulators trained on PPE_1 and PPE_2, the blue curve the R^2 for emulators trained on PPE_3 and PPE_4.

245 3.2.2 Visualization of the parameter-to-output maps

The previously trained emulator can also be used for the visualization of the parameter-to-output dependencies. Such visualizations of the parameter-to-output maps are very useful for informing the user of the effect of a parameter on the outputs. They can guide for instance the sensitivity analysis for the tuning parameters (as we did with our sensitivity analysis based on Sobol indices, shown in Appendix C), and can therefore help selecting the tuning parameters and the corresponding plausible ranges 250 in further tuning exercises.

Here we construct these parameter-to-output maps, similarly to what has been done by Mauritsen et al. (2012), with the important difference that the use of GP emulators in our case allows for a more extensive, or denser, exploration of the selected parameter space. We exemplify such visualizations in Fig. 6, constructed from GP emulators for physics outputs trained on PPE_1 and PPE_2. Every column in the figure corresponds to one parameter being changed, and every row to an output variable. The parameters that are not being changed are kept fixed to their best performing value from PPE_2 (marked with the magenta star in Figures 2 and 3 - although we emphasize that with the trained emulators one can very quickly generate new maps for different parameters). The red shaded areas in each plot denote the allowed output ranges from the observational data. For most of the parameters in PPE_1 varied (all except the $\mathrm{entrpen}$ parameter), the value of global cloud cover (third row of Fig. 6) remains below the lower bound given by the observational data (at 62.7%), which is consistent with our observations in Fig. 2. This is the reason why we selected an increased parameter set $\mathcal{P}_{\mathrm{p2}}$ for the next PPEs, which indeed had a higher influence on the global cloud cover. We refer the reader to Appendix F for the parameter-to-output map constructed from PPE_3 and PPE_4, where the effects of the parameter set $\mathcal{P}_{\mathrm{p2}}$ on global cloud cover can be seen.

The parameters that we identified as mostly influential for the physics output metrics are the critical relative humidity in the upper troposphere (crt) and the coefficient conversion from cloud water to rain (cprcon), influencing the radiation metrics and global cloud cover, together with the coefficient of sedimentation velocity of cloud ice ($\mathrm{cvtfall}$), having a strong non-linear influence on all the physics metrics.

3.3 Tuning of the dynamics outputs

We now discuss the simultaneous tuning of the physics and dynamics outputs. Due to the expected large variability of dynamics outputs which can potentially hinder the training of regression models, we adopt a similar approach to Giorgetta et al. (2018), in that we generate a PPE (PPE_5) and select the best performing model configurations. Also in this case, our criterion for evaluating the model performance gives a higher importance to the global radiation metrics, which are our primary tuning goals, and puts less stringent requirements on the other tuning metrics.

The ML-based tuning of the physics output metrics discussed in the previous section serves as a basis for the second tuning step addressing the dynamics outputs. PPE_5 is generated by simultaneously varying the parameters in the set $\mathcal{P}_{\mathrm{pd}}$ (with LHC sampling), while keeping the other parameters fixed to their best configuration obtained with history matching, from PPE_2 (see Table 5 and the magenta star in Fig. 2 and Fig. 3). The physics parameters in $\mathcal{P}_{\mathrm{pd}}$ are selected based on a sensitivity analysis with Sobol indices, presented in Appendix C. The choice of the dynamics parameters follows Giorgetta et al. (2018), with gkdrag and gkwake chosen for tuning the zonal wind stresses on the ocean surface, and rmscon affecting the zonal mean winds. In Fig. 7 we show the physics (panels (a) and (b)) and the dynamics (panels (c) and (d)) outputs from PPE_5, and highlight the two model configurations (the cyan and the red dot) which achieve the best model performance within PPE_5. The selected configurations are also those closest to the observational range in panel (a) of Fig. 7, given that achieving a balanced TOA radiation has a higher importance in our tuning experiment (Giorgetta et al., 2018). The values of the parameters for these two selected simulations are given in Table 7. These also achieve results comparable with the tuned ICON-aes-1.3, with the TOA radiation balance within the interval $[0,1]$ W/m^2, the TOA long- and short-wave radiation metrics to within 1 W/m^2 from

https://doi.org/10.5194/egusphere-2024-2508
Preprint. Discussion started: 9 August 2024

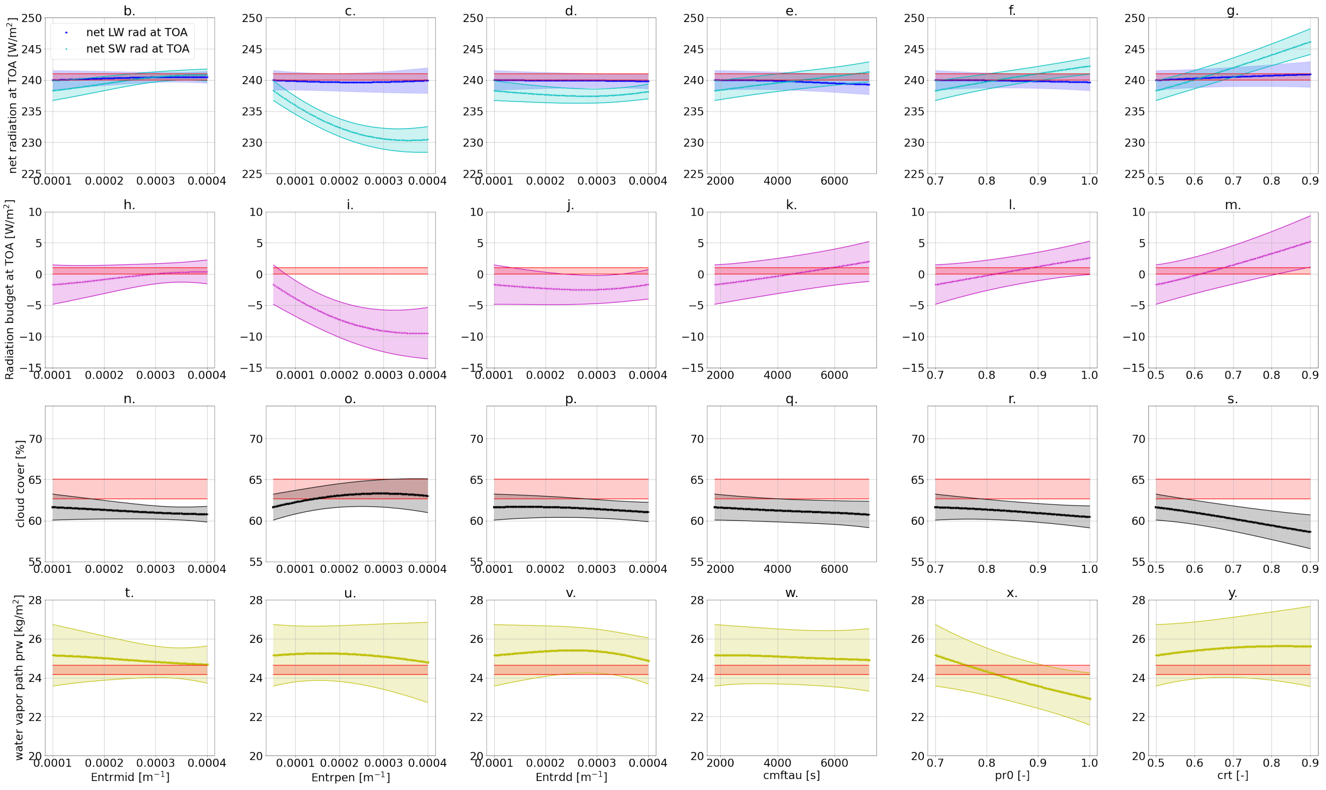


Figure 6. Parameter-to-output maps predicted with the GP-emulators trained on PPE_1 and PPE_2. Every column corresponds to one tuning parameter being changed (see the list in Table 3), and every row to an output metric. The parameters that are not being changed are kept fixed to their best performing value from PPE_2 (marked with the magenta star in Figures 2 and 3). The red shaded areas in each plot denote the allowed output ranges from the observational data. The other colored lines in each plot denote the emulator predictions (for the first row, dark and light blue denote the net long- and short-wave radiation at TOA, respectively), with the corresponding uncertainty (one standard deviation) represented as the shaded area.

the observational range. Also for the other two physics output metrics the performance of the two selected configurations is comparable to ICON-aes-1.3, as they show less than 1% difference in global cloud cover compared to the observational range, and less than 0.5 kg/m^2 difference in the water vapor path. The differences with respect to reference data and ICON-aes-1.3 become more apparent when looking at the dynamics metrics. In panel (c) and (d) it can indeed be seen that the values of these metrics from the reference dataset are not covered by the generated PPE. For most of the metrics the differences of the selected configurations from the reference dataset remain comparable to those of ICON-aes-1.3, except for the mean zonal wind stress over the Southern Ocean (tauu SOO - see panel (c)), where the difference increased from roughly 0.005 N/m^2 to roughly 0.02 N/m^2. The values of the parameters for these two selected runs are given in Table 7. Given the different settings used in the manual tuning for ICON-aes-1.3 (160 km instead of the 80 km resolution used here, and the different time steps used),

https://doi.org/10.5194/egusphere-2024-2508
Preprint. Discussion started: 9 August 2024

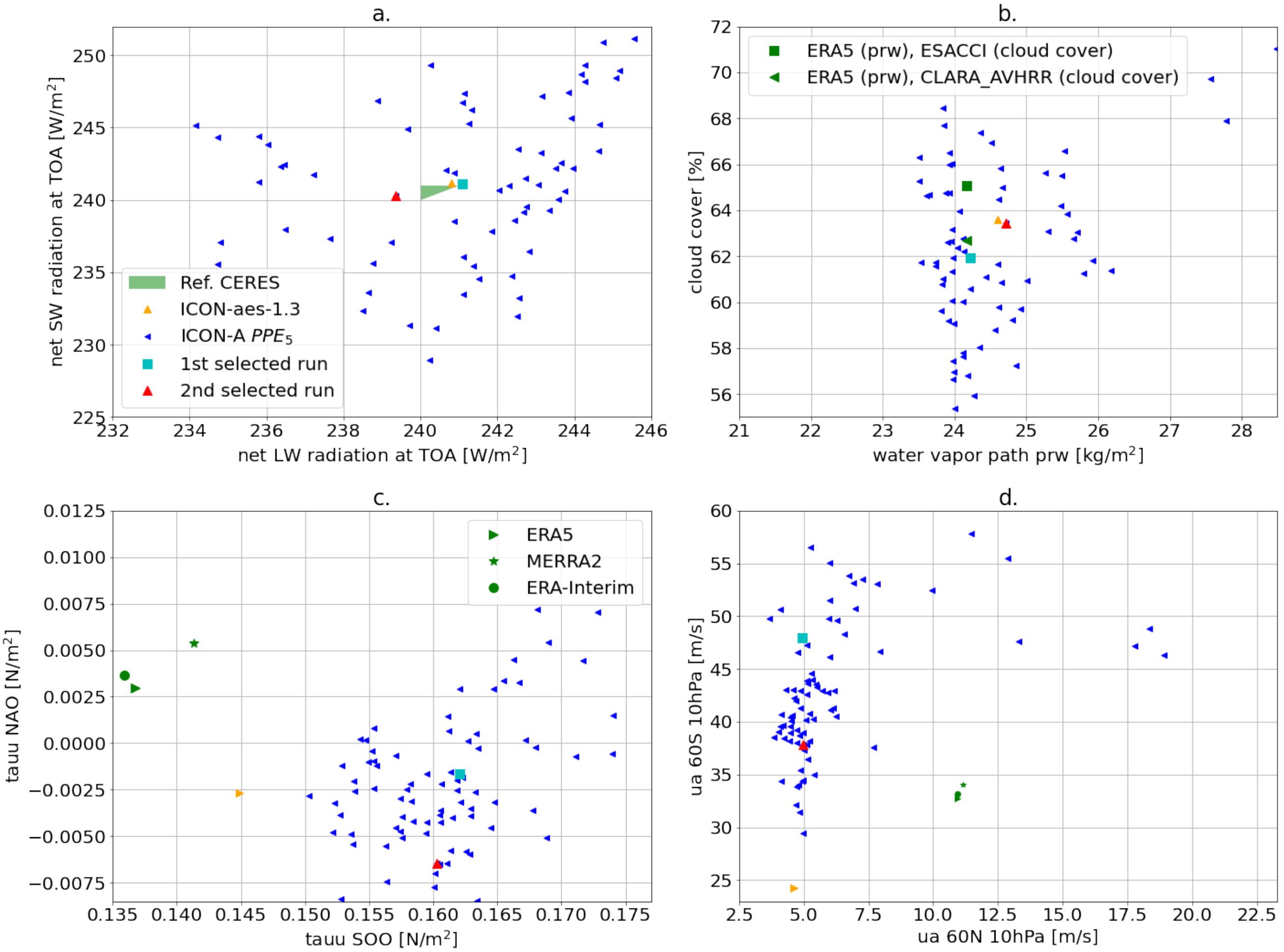


Figure 7. Physics (top row) and dynamics (bottom row) output variables for PPE_5 (blue triangles), compared to ICON-aes-1.3 (orange triangle) and observational datasets. Two selected PPE members corresponding to the best performing configurations are highlighted (cyan square and red triangle).

the differences in the optimal model configurations are not surprising. For instance, the model resolution strongly affects the
295 parameters describing the unresolved orography, and thus the values of the corresponding parameters (Giorgetta et al., 2018).

3.3.1 Analysis of output variability

We now use PPE_5 to analyze the effects of the internal variability of the investigated output metrics and compare them to the parameters' effects. The year-to-year variability of the output metrics is shown in Fig. 8 where we plot the long- vs. short-time averages of the considered outputs, for 30 runs of PPE_5. Additional data complementing the information of Fig. 8 can be found

Physics Parameters	1st selected run	2nd selected run	Giorgetta et al. (2018)
entrmid	2.8526e-4	2.6751e-4	2e-4
entrpen	9.2951e-5	9.2951e-5	2e-4
entrdd	1.7662e-4	1.7662e-4	4e-4
cmftau	2114.6	2114.6	3600
pr0	0.93168	0.93168	1
crt	0.81681	0.80417	0.8
cmfctop	default value: 0.1	default value: 0.1	
cprcon	default value: 2.5e-4	default value: 2.5e-4	
ccsaut	default value: 2.0	default value: 2.0	
csecfrl	default value: 1.5e-5	default value: 1.5e-5	
cvtfall	1.7479	2.00239	
crs	0.88400	0.80222	
csatsc	0.8700	0.64369	
Dynamics Parameters			
gkdrag	0.17404	0.20595	0.1
gkwake	0.08262	0.087592	0.01
rmscon	0.91864	0.82209	0.87
gpicmea	28.375	53.976	
gstd	8.40780	13.025	

Table 7. Values of the parameters for the two members of PPE_5 yielding the best output metrics, shown as cyan square and red triangle in Fig. 7. For comparison, the values of the parameters tuned by Giorgetta et al. (2018) are given as well.

in Appendix E. In Fig. 8 it can be clearly seen that the dynamics outputs (panels in the lower row) have a larger variability across years compared to the physics ones (upper row), which is apparent from the larger spread around the diagonal (no spread would signify no variance), and the larger error bar (which represents the standard deviation over the yearly averages). In each panel we also report the ratio between the mean spread across years S_{yrs} and the PPE spread S_{PPE}, which for each output metric Y are defined as:

$$S_{\mathrm{yrs}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}_{\mathrm{years,i}}(Y)}\,, \tag{6}$$

$$S_{\mathrm{PPE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}\,, \tag{7}$$

where n denotes the size of the PPE, $\mathrm{Var}_{\mathrm{years},i}(Y)$ the variance of output Y over the simulated years for the i-th PPE member, Y_i the ten-year mean of output Y for the i-th PPE member, and $\overline{Y}$ the average of the Y_i over all PPE members. The ratio $S_{\mathrm{years}}/S_{\mathrm{PPE}}$ gives a quantitative measure of the comparison between the yearly output variability and the effects of changing parameters in the PPE. It is clear that for the dynamics outputs, especially the zonal wind stresses on the ocean surface, this ratio is almost one order of magnitude larger than for the physics ones.

An additional source of uncertainty in the dynamics output metrics is their restricted geographical location, which exposes them to biases in spatial patterns. The low variability in the physics variables, which are global means, is consistent with the common observation that already simulations as short as one year can give good tuning results, though using more years, as for instance a full decade used in (Giorgetta et al., 2018), has the benefit to include a larger variation of prescribed boundary conditions as for example El Nino, La Nina or neutral years.

The analysis shown in Fig. 8 shows that for dynamics outputs, the internal variability is almost of the same order of magnitude of the PPE variance, and can therefore partly hide the effects of changing parameters, as discussed above.

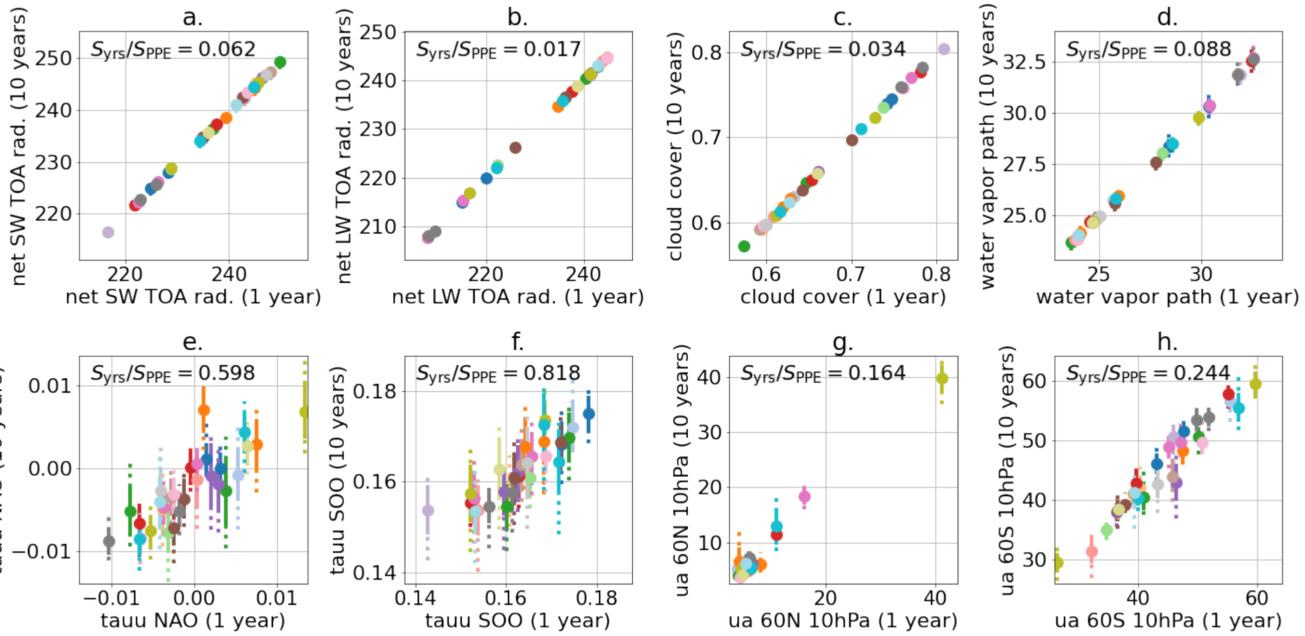


Figure 8. Ten-year mean (1980-1989, y axis) against annual mean (1980, x axis), for the physics (top row, panels (a) to (d)) and dynamics (bottom row, panels (e) to (h)) output variables for 30 runs of PPE_5, represented by different colors. For each data point, the dotted vertical line shows the spread of the annual mean across the ten years (maximum and minimum values), and the solid vertical line denotes one standard deviation, calculated on the 1980-1989 period.

4 Discussions and Conclusions

In this work, we develop an ML-based tuning approach and applied it to the atmospheric component of the ICON climate model (ICON-A). The approach is based on history matching (Williamson et al., 2013, 2017), which balances an extensive exploration of the tuning parameter space with the need of minimizing the number of required ICON-A model simulations. This exploration is aided by an emulator for the outputs of the climate model's PPEs, for which we use Gaussian process (GP) regression. The emulator approximates the climate model simulation outputs and can be used to create large PPEs at a much cheaper computational cost. We first apply history matching to the tuning of physics output metrics (globally averaged radiation and cloud properties), and in a second step we tune also for dynamics output metrics (related to geographically specific atmospheric circulation properties) using a PPE consisting of 80 ten-year ICON-A runs. The ML-based tuning of physics parameterizations with just one iteration of history matching, with a total of 60 model simulations, is already sufficient to achieve a model configuration yielding a global TOA net radiation budget in the range of $[0, 1]$ W/m^2, global radiation metrics and water vapor path consistent with the reference observations, and a globally averaged cloud cover differing by only 2% with respect to the observations.

In the simultaneous PPE-based tuning of physics and dynamics parameterizations, we achieve a TOA radiation balance within the interval $[0, 1]$ W/m^2, TOA long- and short-wave radiation metrics to within 1 W/m^2 from the targeted range, but are not able to reduce the biases in the dynamics output metrics with respect to the previously manually tuned ICON-aes-1.3. The PPE for this tuning step allows us to perform an analysis of the physics and dynamics outputs variability and its comparison with the parameters' effects. This analysis reveals a larger year-to-year variability of the dynamics compared to the physics output metrics. This, combined with the sensitivity of the dynamics metrics to geographic pattern biases, highlights potential limitations that emulator-based approaches may face when tuning for these metrics.

With our presented approach, we have obtained ICON-A model configurations showing an overall performance comparable to ICON-aes-1.3 on most of the selected tuning metrics. Following (Giorgetta et al., 2018), we adopted the two-step approach of first focusing on physics parameters and outputs with one-year long simulations, and then tuning the remaining physics and dynamics parameters and outputs on ten-year long simulations. However, in order to further automize and improve the approach so that all (non-linear) parameter interdependencies and possible feedbacks are properly accounted for, we recommend to move away from this two-step approach for future studies and instead change simultaneously all parameters controlling physics and dynamics outputs in one step. This could be particularly important for tuning coupled models, e.g., for properly accounting for the interactions and feedbacks between atmosphere and ocean. The number of parameters that can be tuned simultaneously is ultimately limited by the available computational resources, since the required size of the PPEs scales linearly with the size of the tuning parameter space. Therefore, sensitivity analysis as presented here becomes a crucial tool to identify and keep only the most important parameters in each model component.

We also note that even though history matching is constructed to minimize the number of climate model simulations for PPEs, this number can still become a limiting factor when tuning models at resolutions higher than the one considered here. Again, including as much prior knowledge as possible in the choice of the parameters, which in a Bayesian setting amounts to the selection of a prior distribution for the optimal parameter values, will be important. Such knowledge of a prior distribution may for instance be obtained by the computationally cheaper tuning of the same model at lower resolutions, provided the same parameterization schemes are used. Incorporating such prior knowledge could reduce the size of the PPEs and the number

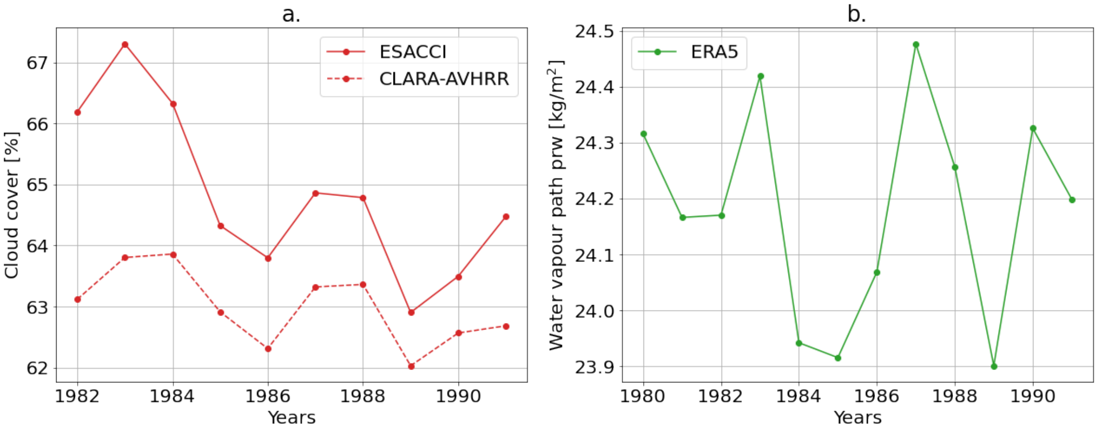


Figure A1. Time series of the observational products used for the cloud cover and the water vapour path

of history matching iterations required to converge to an optimal model configuration, compared to starting from general uninformative priors as we did here (with LHC sampling).

Finally, while here we explored the feasibility of ML-based tuning approaches to improve the tuning of climate models, the seamless integration of such methods within the specific climate modeling framework - to practically enable a largely automatic application - is an aspect that needs to be addressed in further studies. We foresee that incorporating the other tuning steps, such as sensitivity analysis and choice of tuning parameters, their exploration and the evaluation of the outcomes in an automated approach will lead to more accurate and potentially computationally cheaper model tuning, also making this important step in climate model development more objective and reproducible.

Code availability. The code will be published under https://github.com/EyringMLClimateGroup/bonnet24gmd_MLtuning. The software code for the ICON model is available from https://icon-model.org.

Appendix A: Times series of the observational products used

Figure A1 shows the time series of the observational products used for the cloud cover and the water vapour path. The ten year period 1980-1989 was used for the tuning of the dynamic outputs of ICON-A. For the cloud cover observational datasets, the earliest year available is 1982, therefore we added the years 1990-1991 in our tuning analysis. The variability in the years illustrates the internal climate variability. We remark that other observational products exist for these outputs but do not include the studied years. For example, ESACCI-WATERVAPOUR starts from year 2002, MODIS starts from year 2002, or Cloudsat starts from year 2006.

https://doi.org/10.5194/egusphere-2024-2508
Preprint. Discussion started: 9 August 2024

Appendix B: Details on GP emulators and choice of the underlying hyperparameters

In this appendix we give a brief description of the Gaussian process (GP) regression framework used to construct emulators in this work, and provide the relevant details regarding the hyperparameters used in their implementation. Gaussian processes are widely used in the context of Bayesian optimization, as they are a method for describing distributions over unknown functions, and can be efficiently updated, or trained, using samples from the ground-truth distribution (Rasmussen and Williams, 2005). In our case, the function we want to approximate with GP regression is that describing the dependence of a specific output Y of the climate model, on a set of tuning parameters $\boldsymbol{x}$, which we call $Y_{\text{model}}(\boldsymbol{x})$. The output of a Gaussian process trained on set $\mathcal{T} = \{\boldsymbol{x}_i, Y_{\text{model}}(\boldsymbol{x}_i)\}_i$ of ground-truth samples (ICON-A model runs in our case) can be written as:

$$f(\boldsymbol{x})\,|\,\mathcal{T} \sim \mathcal{GP}(\mu(\cdot), K_{\cdot,\cdot})\,, \tag{B1}$$

where $\mathcal{GP}$ denotes the GP function distribution with $\mu(\boldsymbol{x})$ and K respectively being the mean function and the covariance matrix that implicitly depend on $\mathcal{T}$, i.e., have been updated with the knowledge of the training data $\mathcal{T}$ using Bayes' rule. Closed form expressions for these functions are available and can be found in Rasmussen and Williams (2005). That is to say, given a new configuration $\boldsymbol{x}$ of tuning parameters, a GP trained on an ICON-A PPE for a given variable Y would output a normally distributed random variable with mean $\mu(\boldsymbol{x})$ and variance $\sigma^2(\boldsymbol{x})$ (which can also be explicitly calculated from the knowledge of the covariance matrix K (Rasmussen and Williams, 2005)). We therefore interpret $\mu(\boldsymbol{x})$ as our GP emulator prediction for Y, and $\sigma^2(\boldsymbol{x})$ as the associated uncertainty, and write:

$$Y_{\text{emul}}(\boldsymbol{x}) \equiv \mu(\boldsymbol{x})\,, \tag{B2}$$

$$\text{Var}(Y_{\text{emul}}(\boldsymbol{x})) \equiv \sigma^2(\boldsymbol{x})\,, \tag{B3}$$

which we use in Eq. (1) in the main text.

Importantly, the properties of the GP, in particular of the covariance matrix K, depend on the choice of a *kernel function* $k(\boldsymbol{x}, \boldsymbol{x'})$, which describes how the predictions at two points $\boldsymbol{x}$ and $\boldsymbol{x'}$ are correlated. Kernel functions may also contain trainable hyperparameters, which are typically optimized by maximizing the log marginal likelihood with respect to the training dataset (Rasmussen and Williams, 2005).

For our implementations we used the GP regression library implemented in scikit-learn package (https://scikit-learn.org/stable/). We found Matèrn kernels to yield the highest prediction accuracy (which we measure via R^2 coefficient). Matèrn kernels have two hyperparameters: a lengthscale l and a smoothness parameter ν. The length scale is typically the distance by which one can extrapolate outside the training data points: smaller values of l correspond to more rapidly varying functions that the GP can fit. This hyperparameter, together with the overall scale of the kernel, is optimized using the L-BFGS-B optimization (Jorge Nocedal, 2006) pre-implemented in scikit-learn. For the smoothness parameter ν, four values were tested: $\nu = 0.5$ corresponds to the absolute exponential kernel, $\nu = 1.5$ to a one-time differentiable function, $\nu = 2.5$ to a twice differentiable function and $\nu \to \infty$ to a radial basis function (RBF) kernel. These four values of ν allow a computational cost around ten times smaller than other values, since they do not require to evaluate the modified Bessel function (Rasmussen, 2006). The

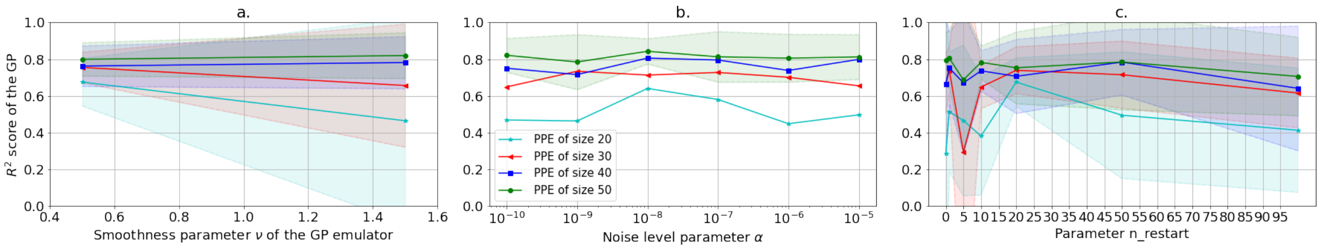


Figure B1. Performance (R^2 coefficient, calculated with 5-fold cross-validation) of the GP emulator with Matèrn kernel trained on PPE_1 and PPE_2, for different choices of hyperparameters: (a) for different values of ν, (b) for different values of α, and (c) for different values of n_restart.

values of $\nu = 2.5$ and $\nu \to \infty$ yield large negative R^2-scores, so are not represented here. In panel (a) of Fig. B1 we observe a comparable performance of the GP emulator for $\nu = 0.5$ (absolute exponential kernel) and $\nu = 1.5$.

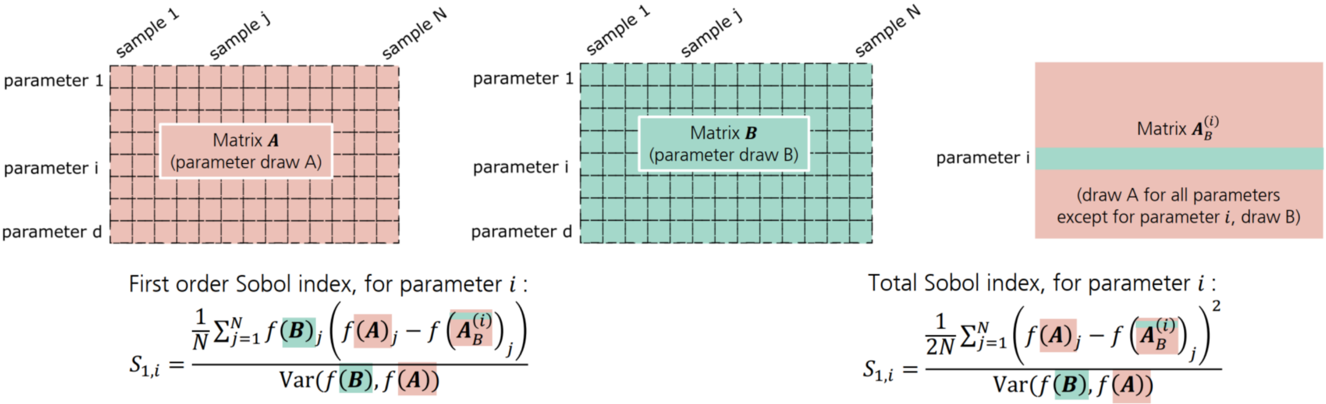
Other hyperparameters in the GP optimization are the noise level α (which can be interpreted as the variance of Gaussian noise added to the training data, with the aim of increasing the numerical stability of GP evaluations) and the number of random hyperparameter initializations for the log marginal likelihood optimization (denoted with n_restart). Several values of α between 10^{-15} and 10^{-5} were tested. We show these tests in panel (b) Fig. B1. The values of $\alpha < 10^{-10}$ yield large negative R^2 scores. A change of α for $10^{-10} < \alpha < 10^{-5}$ does not have a significant effect on the performance of the GP emulator. Finally, we also tested several values of the n_restart, between 0 and 100, as shown in panel (c) of Fig. B1. From the tests presented in Fig. B1, the following values of the three hyperparameters are chosen (which are also default values in scikit-learn): $\nu = 1.5$, $\alpha = 10^{-10}$ and n_restart$= 0$.

Appendix C: Sobol indices for the physics parameters and outputs

The first order Sobol indices were calculated using the formula (b) in Table 2 of (Saltelli et al., 2010) normalized by the variance $\mathrm{Var}(f(\boldsymbol{A}), f(\boldsymbol{B}))$ and the total order Sobol indices were calculated using the formula (f) in Table 2 of (Saltelli et al., 2010) (see Figure C1 for a schematic representation of the matrices and the formulas), with $N = 70000$ (size of the samples used in the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$) allowing a converged value of the indices. Here f refers to the prediction of GP emulator for one output metric calculated for a given parameter sampling (arranged in a matrix e.g., $\boldsymbol{A}$).

The higher the values of the first and total Sobol indices for a parameter and corresponding output, the higher the influence of that parameter on that output. Figure C2 shows estimators of the first and total Sobol indices for all physics parameters and physics outputs. The parameter crt has a high influence on the radiation metrics (Fig. C2(a), Fig. C2(c) and Fig. C2(h)) and the cloud cover (Fig. C2(d)). The parameter $\mathrm{cvtfall}$ has a high influence on the net LW radiation at the TOA (Fig. C2(g)) and on the cloud cover (Fig. C2(i)). For the choice of physics parameters for PPE_5, these two parameters were kept. In addition, two new physics tuning parameters were added following expert advice: crs and csatsc.

https://doi.org/10.5194/egusphere-2024-2508
Preprint. Discussion started: 9 August 2024



$$S_{1,i} = \frac{\frac{1}{N}\Sigma_{j=1}^{N} f(\boldsymbol{B})_j \left(f(\boldsymbol{A})_j - f\left(\boldsymbol{A}_B^{(i)}\right)_j\right)}{\mathrm{Var}(f(\boldsymbol{B}), f(\boldsymbol{A}))}$$

$$S_{1,i} = \frac{\frac{1}{2N}\Sigma_{j=1}^{N} \left(f(\boldsymbol{A})_j - f\left(\boldsymbol{A}_B^{(i)}\right)_j\right)^2}{\mathrm{Var}(f(\boldsymbol{B}), f(\boldsymbol{A}))}$$

Figure C1. Schematic representation of Sobol indices and formulas of the first order and total Sobol indices.

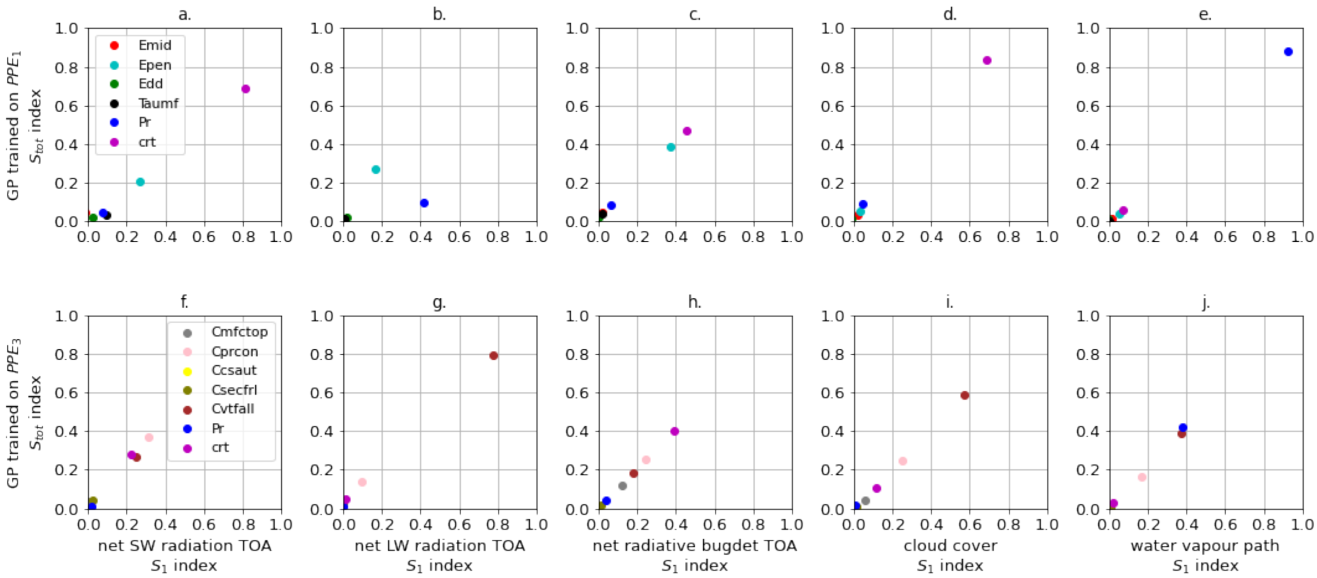


Figure C2. First order Sobol index S_1 (x-axis) and total Sobol index S_{tot} (y-axis) for the physics parameters (in legend) and outputs, net SW radiation at the TOA (panels (a) and (f)), net LW radiation at TOA (panels (b) and (g)), net radiative bugdet at TOA (panels (c) and (h)), cloud cover (panels (d) and (i)), water vapour path (panels (e) and (j)) using the GP trained on PPE_1 for panels (a) to (e) and trained on PPE_3 for panels (f) to (j).

Appendix D: Additional information on the generated PPEs

In this appendix we show additional data for the PPEs we generated in this work. Specifically, in Fig. D1 we show the sampled parameter values for PPE_3 (red circles) and PPE_4 (grey triangles), where signs of (slow) convergence of history matching are visible already after one iteration (in the distribution of the members of PPE_4 being slightly shifted and narrower). Figure D2 shows the sampled parameter values for PPE_5 (blue triangles), with the cyan square and red triangle marking the best performing configurations reported in Table 7 in the main text.

Appendix E: Times series of the physics and dynamic metrics

In this appendix we show additional information complementing Fig. 8 in Section 3.2.1 in the main text. In Fig. E1 we show the yearly averages of the physics (top row, panels (a) to (d)) and dynamics (bottom row, panels (e) to (h)) output variables for the 30 runs of PPE_5 corresponding to Fig. 8. Also in these time series the higher year-to-year variability of the dynamics outputs compared to the physics ones can be clearly seen.

Appendix F: Additional information on parameter-to-output maps

In this appendix we show additional information on the parameter-to-output maps discussed in Section 3.2.2. In particular, in Fig. F1 we show the parameter-to-output map predicted with the GP-emulators trained on PPE_3 and PPE_4, on parameter set $\mathcal{P}_{\mathrm{p2}}$. Here we can see that parameter set $\mathcal{P}_{\mathrm{p2}}$ does indeed allow for a higher (and closer to the observational values) global cloud cover compared to $\mathcal{P}_{\mathrm{p1}}$ (see Fig. 6).

Author contributions. PB developed the ML-based tuning approach and performed the PPE simulations with ICON-A. She led the analysis with support of LP. VE formulated the research question and concept. All authors discussed the methodology and findings. LP and PB wrote the manuscript with contributions from all authors.

Competing interests. The authors declare no competing interests.

Acknowledgements. This study was funded by the European Research Council (ERC) Synergy Grant “Understanding and Modeling the Earth System with Machine Learning“ (USMILE) under the Horizon 2020 Research and Innovation program (Grant Agreement No. 855187) and by the “Earth System Models for the Future (ESM2025)“ project under the European Union’s Horizon 2020 research and innovation programme (Grant Agreement No. 101003536). VE was additionally supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the Gottfried Wilhelm Leibniz Prize awarded to Veronika Eyring (Reference No. EY 22/2-1). The contribution of LP and MS was made possible by the DLR Quantum Computing Initiative and the Federal Ministry for Economic Affairs and Climate Action; https:

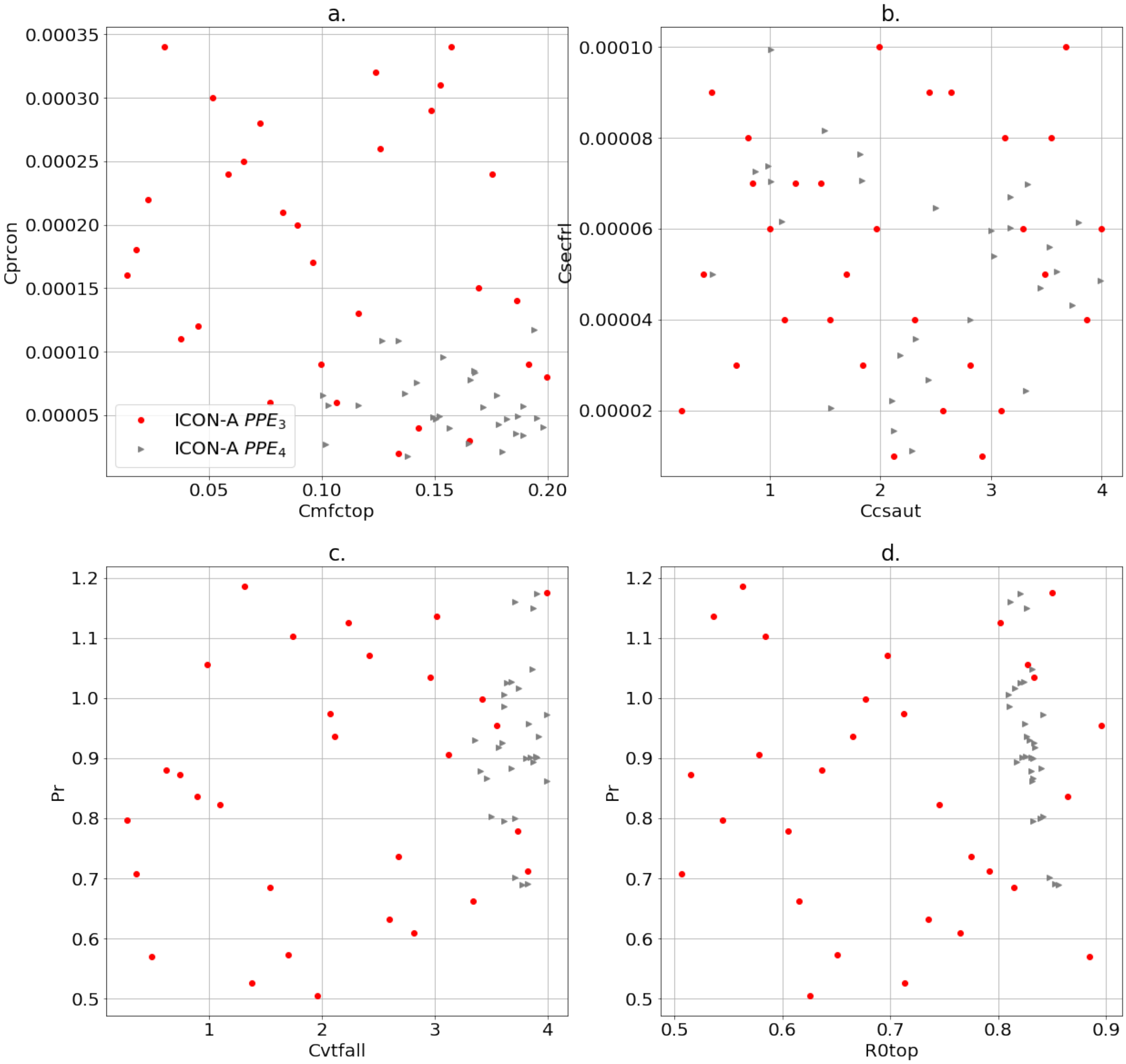


Figure D1. Sampled parameter values for PPE_3 (red circles) and PPE_4 (grey triangles). For each panel, two parameters are plotted on the two axes (see Table 3). The two PPEs are generated with parameter set $\mathcal{P}_{\mathrm{p2}}$. Signs of (slow) convergence of history matching are visible already after one iteration (in the distribution of the members of PPE_4 being slightly shifted and narrower).

//qci.dlr.de/projects/klim-qml. This work used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID 1179 (USMILE) and 1083 (Climate Informatics).

References

Cleary, E., Garbuno-Inigo, A., Lan, S., Schneider, T., and Stuart, A. M.: Calibrate, emulate, sample, Journal of Computational Physics, 424, 109 716, https://doi.org/10.1016/j.jcp.2020.109716, 2021.

Couvreux, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque, N., Rio, C., Audouin, O., Salter, J., Bazile, E., Brient, F., Favot, F., Honnert, R., Lefebvre, M.-P., Madeleine, J.-B., Rodier, Q., and Xu, W.: Process-Based Climate Model Development Harnessing Machine Learning: I. A Calibration Tool for Parameterization Improvement, Journal of Advances in Modeling Earth Systems, 13, https://doi.org/10.1029/2020ms002217, 2021.

Crueger, T., Giorgetta, M. A., Brokopf, R., Esch, M., Fiedler, S., Hohenegger, C., Kornblueh, L., Mauritsen, T., Nam, C., Naumann, A. K., Peters, K., Rast, S., Roeckner, E., Sakradzija, M., Schmidt, H., Vial, J., Vogel, R., and Stevens, B.: ICON-A, The Atmosphere Component of the ICON Earth System Model: II. Model Evaluation, Journal of Advances in Modeling Earth Systems, 10, 1638–1662, https://doi.org/10.1029/2017ms001233, 2018.

Dagon, K., Sanderson, B. M., Fisher, R. A., and Lawrence, D. M.: A machine learning approach to emulation and biophysical parameter estimation with the Community Land Model, version 5, Advances in Statistical Climatology, Meteorology and Oceanography, 6, 223–244, https://doi.org/10.5194/ascmo-6-223-2020, 2020.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J., Park, B., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137, 553–597, https://doi.org/10.1002/qj.828, 2011.

Dunbar, O. R. A., Garbuno-Inigo, A., Schneider, T., and Stuart, A. M.: Calibration and Uncertainty Quantification of Convective Parameters in an Idealized GCM, Journal of Advances in Modeling Earth Systems, 13, https://doi.org/10.1029/2020ms002454, 2021.

Eidhammer, T., Gettelman, A., Thayer-Calder, K., Watson-Parris, D., Elsaesser, G., Morrison, H., van Lier-Walqui, M., Song, C., and McCoy, D.: An Extensible Perturbed Parameter Ensemble (PPE) for the Community Atmosphere Model Version 6, EGUsphere, 2024, 1–27, https://doi.org/10.5194/egusphere-2023-2165, 2024.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B.: The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), Journal of Climate, 30, 5419–5454, https://doi.org/10.1175/jcli-d-16-0758.1, 2017.

Gentine, P., Eyring, V., and Beucler, T.: Deep Learning for the Parametrization of Subgrid Processes in Climate Models, chap. 21, pp. 307–314, John Wiley & Sons, Ltd, Chichester, West Sussex, UK, https://doi.org/https://doi.org/10.1002/9781119646181.ch21, 2021.

Giorgetta, M. A., Brokopf, R., Crueger, T., Esch, M., Fiedler, S., Helmert, J., Hohenegger, C., Kornblueh, L., Köhler, M., Manzini, E., Mauritsen, T., Nam, C., Raddatz, T., Rast, S., Reinert, D., Sakradzija, M., Schmidt, H., Schneck, R., Schnur, R., Silvers, L., Wan, H., Zängl, G., and Stevens, B.: ICON-A, the Atmosphere Component of the ICON Earth System Model: I. Model Description, Journal of Advances in Modeling Earth Systems, 10, 1613–1637, https://doi.org/10.1029/2017ms001242, 2018.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G. D., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., Folini, D., Ji, D., Klocke, D., Qian, Y., Rauser, F., Rio, C., Tomassini, L., Watanabe, M., and Williamson, D.: The Art and Science of Climate Model Tuning, Bulletin of the American Meteorological Society, 98, 589–602, https://doi.org/10.1175/bams-d-15-00135.1, 2017.

Hourdin, F., Williamson, D., Rio, C., Couvreux, F., Roehrig, R., Villefranque, N., Musat, I., Fairhead, L., Diallo, F. B., and Volodina, V.: Process-Based Climate Model Development Harnessing Machine Learning: II. Model Calibration From Single Column to Global, Journal of Advances in Modeling Earth Systems, 13, https://doi.org/10.1029/2020ms002225, 2021.

Hourdin, F., Ferster, B., Deshayes, J., Mignot, J., Musat, I., and Williamson, D.: Toward machine-assisted tuning avoiding the underestimation of uncertainty in climate change projections, Science Advances, 9, https://doi.org/10.1126/sciadv.adf2758, 2023.

ICON: ICON: Icosahedral Nonhydrostatic Weather and Climate Model, https://code.mpimet.mpg.de/projects/iconpublic, 2015.

Iturbide, M., Gutiérrez, J. M., Alves, L. M., Bedia, J., Cerezo-Mota, R., Cimadevilla, E., Cofiño, A. S., Luca, A. D., Faria, S. H., Gorodetskaya, I. V., Hauser, M., Herrera, S., Hennessy, K., Hewitt, H. T., Jones, R. G., Krakovska, S., Manzanas, R., Martínez-Castro, D., Narisma, G. T., Nurhati, I. S., Pinto, I., Seneviratne, S. I., van den Hurk, B., and Vera, C. S.: An update of IPCC climate reference regions for subcontinental analysis of climate model data: definition and aggregated datasets, Earth System Science Data, 12, 2959–2970, https://doi.org/10.5194/essd-12-2959-2020, 2020.

Jorge Nocedal, S. J. W.: Numerical Optimization, Springer New York, https://doi.org/10.1007/978-0-387-40065-5, 2006.

Jungclaus, J. H., Lorenz, S. J., Schmidt, H., Brovkin, V., Brüggemann, N., Chegini, F., Crüger, T., De-Vrese, P., Gayler, V., Giorgetta, M. A., Gutjahr, O., Haak, H., Hagemann, S., Hanke, M., Ilyina, T., Korn, P., Kröger, J., Linardakis, L., Mehlmann, C., Mikolajewicz, U., Müller, W. A., Nabel, J. E. M. S., Notz, D., Pohlmann, H., Putrasahan, D. A., Raddatz, T., Ramme, L., Redler, R., Reick, C. H., Riddick, T., Sam, T., Schneck, R., Schnur, R., Schupfner, M., Storch, J.-S., Wachsmann, F., Wieners, K.-H., Ziemen, F., Stevens, B., Marotzke, J., and Claussen, M.: The ICON Earth System Model Version 1.0, Journal of Advances in Modeling Earth Systems, 14, https://doi.org/10.1029/2021ms002813, 2022.

Karlsson, K.-G., Anttila, K., Trentmann, J., Stengel, M., Solodovnik, I., Meirink, J. F., Devasthale, A., Kothe, S., Jääskeläinen, E., Sedlar, J., Benas, N., van Zadelhoff, G.-J., Stein, D., Finkensieper, S., Håkansson, N., Hollmann, R., Kaiser, J., and Werscheck, M.: CLARA-A2.1: CM SAF cLoud, Albedo and surface RAdiation dataset from AVHRR data - Edition 2.1, https://doi.org/10.5676/EUM_SAF_CM/CLARA_AVHRR/V002_01, 2020.

Kato, S., Loeb, N. G., Rose, F. G., Doelling, D. R., Rutan, D. A., Caldwell, T. E., Yu, L., and Weller, R. A.: Surface Irradiances Consistent with CERES-Derived Top-of-Atmosphere Shortwave and Longwave Irradiances, Journal of Climate, 26, 2719–2740, https://doi.org/10.1175/jcli-d-12-00436.1, 2013.

Loeb, N. G., Wielicki, B. A., Doelling, D. R., Smith, G. L., Keyes, D. F., Kato, S., Manalo-Smith, N., and Wong, T.: Toward Optimal Closure of the Earth's Top-of-Atmosphere Radiation Budget, Journal of Climate, 22, 748–766, https://doi.org/10.1175/2008jcli2637.1, 2009.

Loeppky, J. L., Sacks, J., and Welch, W. J.: Choosing the Sample Size of a Computer Experiment: A Practical Guide, Technometrics, 51, 366–376, https://doi.org/10.1198/tech.2009.08040, 2009.

Mansfield, L. A. and Sheshadri, A.: Calibration and Uncertainty Quantification of a Gravity Wave Parameterization: A Case Study of the Quasi-Biennial Oscillation in an Intermediate Complexity Climate Model, Journal of Advances in Modeling Earth Systems, 14, https://doi.org/10.1029/2022ms003245, 2022.

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., Haak, H., Jungclaus, J., Klocke, D., Matei, D., Mikolajewicz, U., Notz, D., Pincus, R., Schmidt, H., and Tomassini, L.: Tuning the climate of a global model, Journal of Advances in Modeling Earth Systems, 4, n/a–n/a, https://doi.org/10.1029/2012ms000154, 2012.

Mignot, J., Hourdin, F., Deshayes, J., Boucher, O., Gastineau, G., Musat, I., Vancoppenolle, M., Servonnat, J., Caubel, A., Chéruy, F., Denvil, S., Dufresne, J.-L., Ethé, C., Fairhead, L., Foujols, M.-A., Grandpeix, J.-Y., Levavasseur, G., Marti, O., Menary, M., Rio, C., Rousset, C., and Silvy, Y.: The Tuning Strategy of IPSL-CM6A-LR, Journal of Advances in Modeling Earth Systems, 13, https://doi.org/10.1029/2020ms002340, 2021.

NASA/LARC/SD/ASDC: CERES Energy Balanced and Filled (EBAF) TOA and Surface Monthly means data in netCDF Edition 4.1, https://doi.org/10.5067/TERRA-AQUA/CERES/EBAF_L3B.004.1, 2019.

Rasmussen, C. E.: Gaussian processes for machine learning, MIT Press, https://gaussianprocess.org/gpml/chapters/RW.pdf, 2006.

Rasmussen, C. E. and Williams, C. K. I.: Gaussian Processes for Machine Learning, The MIT Press, https://doi.org/10.7551/mitpress/3206.001.0001, 2005.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S.: Global Sensitivity Analysis. The Primer, Wiley, https://doi.org/10.1002/9780470725184, 2007.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S.: Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index, Computer Physics Communications, 181, 259–270, https://doi.org/10.1016/j.cpc.2009.09.018, 2010.

Schmidt, G. A., Bader, D., Donner, L. J., Elsaesser, G. S., Golaz, J.-C., Hannay, C., Molod, A., Neale, R. B., and Saha, S.: Practice and philosophy of climate model tuning across six US modeling centers, Geoscientific Model Development, 10, 3207–3223, https://doi.org/10.5194/gmd-10-3207-2017, 2017.

Stengel, M., Stapelberg, S., Sus, O., Schlundt, C., Poulsen, C., Thomas, G., Christensen, M., Carbajal Henken, C., Preusker, R., Fischer, J., Devasthale, A., Willén, U., Karlsson, K.-G., McGarragh, G. R., Proud, S., Povey, A. C., Grainger, R. G., Meirink, J. F., Feofilov, A., Bennartz, R., Bojanowski, J. S., and Hollmann, R.: Cloud property datasets retrieved from AVHRR, MODIS, AATSR and MERIS in the framework of the Cloud cci project, Earth System Science Data, 9, 881–904, https://doi.org/10.5194/essd-9-881-2017, 2017.

Tebaldi, C., Debeire, K., Eyring, V., Fischer, E., Fyfe, J., Friedlingstein, P., Knutti, R., Lowe, J., O'Neill, B., Sanderson, B., van Vuuren, D., Riahi, K., Meinshausen, M., Nicholls, Z., Tokarska, K. B., Hurtt, G., Kriegler, E., Lamarque, J.-F., Meehl, G., Moss, R., Bauer, S. E., Boucher, O., Brovkin, V., Byun, Y.-H., Dix, M., Gualdi, S., Guo, H., John, J. G., Kharin, S., Kim, Y., Koshiro, T., Ma, L., Olivié, D., Panickal, S., Qiao, F., Rong, X., Rosenbloom, N., Schupfner, M., Séférian, R., Sellar, A., Semmler, T., Shi, X., Song, Z., Steger, C., Stouffer, R., Swart, N., Tachiiri, K., Tang, Q., Tatebe, H., Voldoire, A., Volodin, E., Wyser, K., Xin, X., Yang, S., Yu, Y., and Ziehn, T.: Climate model projections from the Scenario Model Intercomparison Project (ScenarioMIP) of CMIP6, Earth System Dynamics, 12, 253–293, https://doi.org/10.5194/esd-12-253-2021, 2021.

Watson-Parris, D., Williams, A., Deaconu, L., and Stier, P.: Model calibration using ESEm v1.1.0 – an open, scalable Earth system emulator, Geoscientific Model Development, 14, 7659–7672, https://doi.org/10.5194/gmd-14-7659-2021, 2021.

565 Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., and Yamazaki, K.: History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble, Climate Dynamics, 41, 1703–1729, https://doi.org/10.1007/s00382-013-1896-4, 2013.

Williamson, D. B., Blaker, A. T., and Sinha, B.: Tuning without over-tuning: parametric uncertainty quantification for the NEMO ocean model, Geoscientific Model Development, 10, 1789–1816, https://doi.org/10.5194/gmd-10-1789-2017, 2017.

570 Zhang, T., Li, L., Lin, Y., Xue, W., Xie, F., Xu, H., and Huang, X.: An automatic and effective parameter optimization method for model tuning, Geoscientific Model Development, 8, 3579–3591, https://doi.org/10.5194/gmd-8-3579-2015, 2015.

Zängl, G., Reinert, D., Rípodas, P., and Baldauf, M.: The ICON (ICOsahedral Non-hydrostatic) modelling frameworkof DWD and MPI-M: Description of the non-hydrostaticdynamical core, Quarterly Journal of the Royal Meteorological Society, 141, 563–579, https://doi.org/10.1002/qj.2378, 2014.

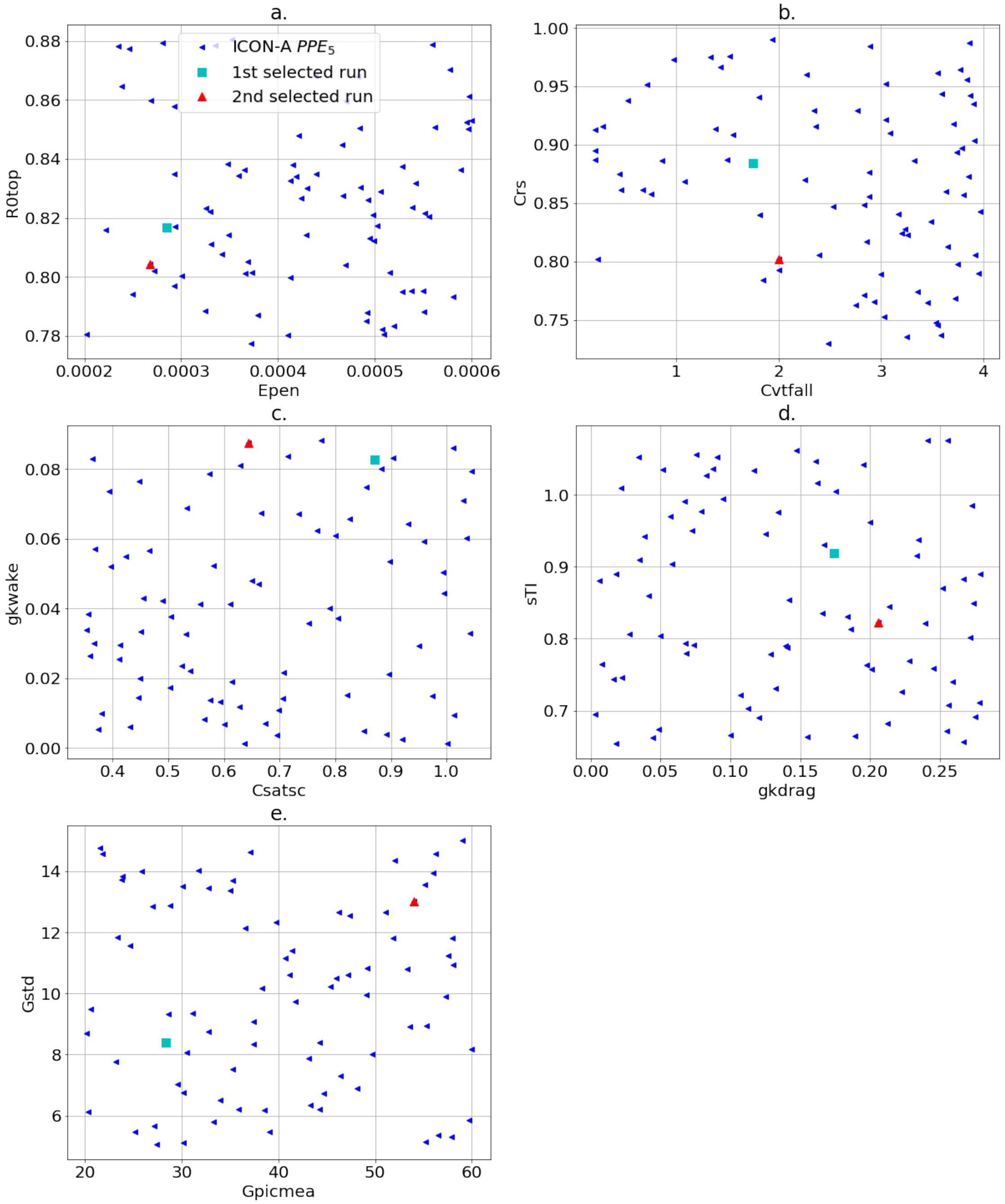


Figure D2. Sampled parameter values for PPE_5 (blue triangles). For each panel, two parameters are plotted on the two axes (see Tables 3 and 4). The PPE is generated with parameter set $\mathcal{P}_{\mathrm{pd}}$. Two selected PPE members corresponding to the best performing configurations are highlighted (cyan square and red triangle).

https://doi.org/10.5194/egusphere-2024-2508
Preprint. Discussion started: 9 August 2024

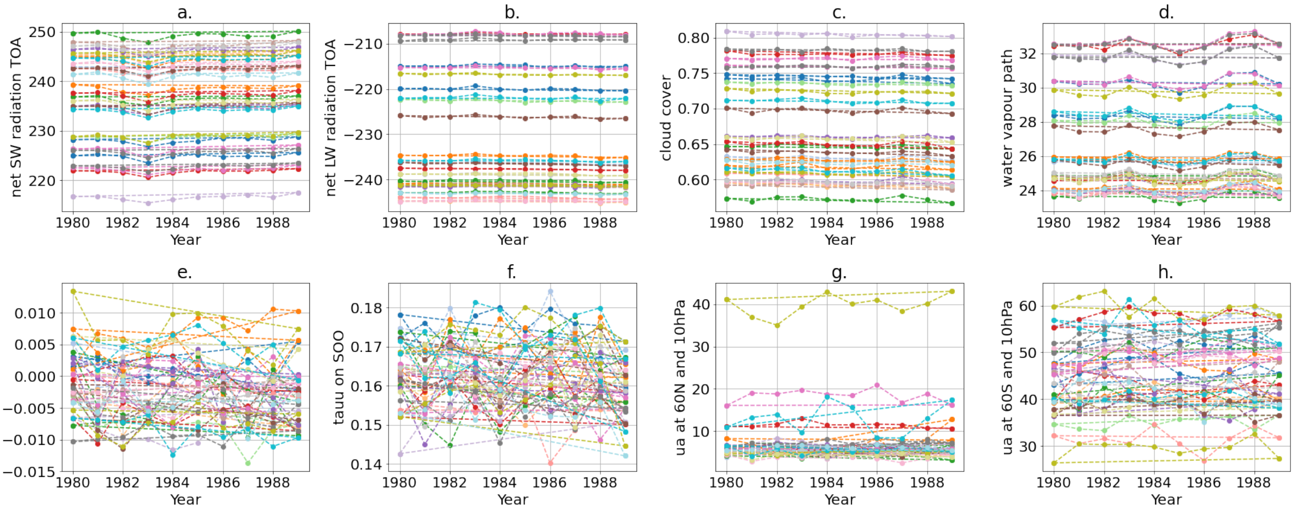


Figure E1. Time series (yearly averages) of the physics (top row, panels (a) to (d)) and dynamics (bottom row, panels (e) to (h)) output variables for 30 runs of PPE_5 (each color corresponds to one run). The values at year 1980 and 1989 are connected with a dashed line to help the reader identify the runs.

https://doi.org/10.5194/egusphere-2024-2508
Preprint. Discussion started: 9 August 2024

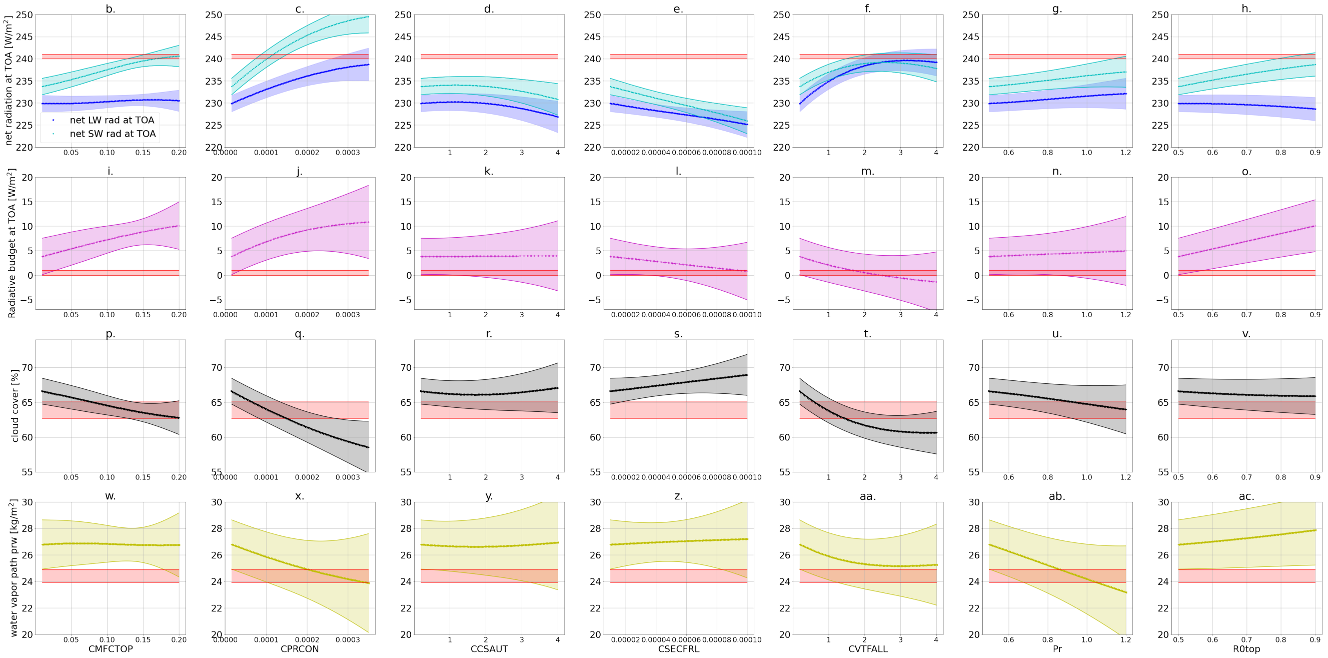


Figure F1. Parameter-to-output map predicted with the GP-emulators trained on PPE_3 and PPE_4. Every column corresponds to one tuning parameter being changed (see the list in Table 3), and every row to to an output variable. The parameters that are not being changed are kept fixed to their best performing value from PPE_2 (marked with the magenta star in Figures 2 and 3). The red shaded areas in each plot denote the allowed output ranges from the observational data. The other colored lines in each plot denote the emulator predictions (for the first row, dark and light blue denote the net long- and short-wave radiation at TOA, respectively), with the corresponding uncertainty (one standard deviation) represented as the shaded area.