Response to reviewer #2 on the manuscript: **Linear Meta-Model optimization for regional climate models (LiMMo version 1.0)**

by Sergei Petrov and Beate Geyer

July 9, 2025

Dear Anonymous reviewer #2, thank you very much for your valuable comments. You will find detailed answers to your comments in the following text. We will refer to the initially submitted version of the manuscript with a red background color and to the revised version with a green background color. We will use a yellow background color for your comments.

First of all, I thank the authors for this articles. I think its core concept and developments are valuable, interesting and worth publishing. I had a great time reading it.

We appreciate this comment and thank the reviewer for their critical and constructive feedback. We hope the revised manuscript meets GMD's high scientific standards, and we hope the reviewer has a similar experience with our replies and revision as with the original manuscript.

1. My biggest comment on the content is the following. For the study to be complete, I believe there lacks a comparison with another method (e.g., quadratic regression) which could be expected to result in a better set of parameters at the expense of more time and computing resources. I do not request adding this comparison to the study. If you do, I think it would be extremely valuable (I know that your goal is to be cost-efficient, but the method is anyways, yet here we are talking about presenting it for the first time, which could deserve a one-time investment in order to more clearly situate its pros and cons with respects to other methods), but if you don't, please at least develop on the potential problems brought by the linear approximation. I think that the current manuscript version is very superficial on that point, highlighting the cost advantage but overlooking the cons.

We agree with the reviewer that a comparison with a quadratic method would be valuable. However, this will have to wait for future work, as it requires substantial additional resources. Nevertheless, we believe that our results are worth publishing, even without a direct comparison to an alternative configuration optimization method. The minimum number of simulations required for quadratic regression training is $1 + 2 \cdot N + \frac{N \cdot (N-1)}{2}$ for N parameters. For a set of 12 parameters, this would require 91 five-year simulations. Even with the 13 simulations we have already conducted, we would still need 78 more. Running $78 \times 5 = 390$ simulation years is common in climate change studies, but it exceeds the resources available for this work. Future studies probably could carefully compare fewer parameters. In the meantime, we will partially close the gap by adding Fig. 1 to the manuscript (Fig.7 in the revised version).

A similar plot was produced by Bellprat et al. 2012 (see Fig. 2), where quadratic regression was implemented. Of course, the optimization setups differ. The main difference is that in Bellprat et al. 2012 COSMO-CLM (Rockel, Will, and Hense 2008) was optimized at a spatial resolution of 50 km, whereas we use ICON-CLM at a spatial resolution of 12 km. The quadratic regression was trained for on only three variables — 2-m temperature, daily precipitation, and total cloud cover (we used shortwave radiation flux instead and three additional quantities: **tasmin**, **tasmax** and **hfls**). In Bellprat et al. 2012 regression yields monthly mean values over five years (60 values per grid point per variable), whereas in our approach regression yields 12 five-year mean monthly averages (i.e., the average of January, February, etc. from 2003 to 2007). The set of parameters considered is also slightly different.

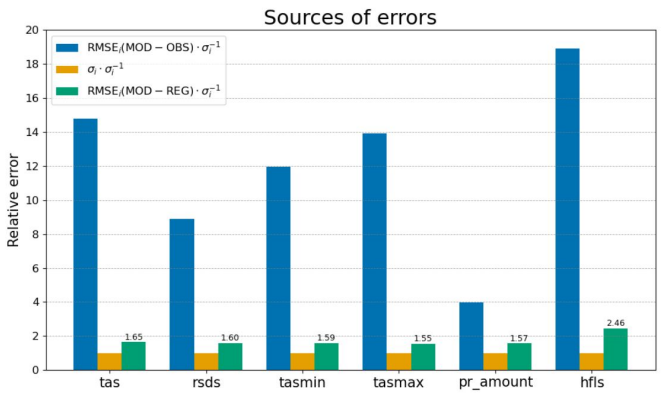

Figure 1: The comparison of the different sources of error in LiMMo. Values are normalized on the intrinsic variability of the ICON-CLM (Eq.2) for each model variable. The blue bar shows the RMSE of the ICON-CLM output with the NWP configuration to the observations. The orange bar shows the intrinsic variability (Eq.2). The green bar shows the RMSE between the ICON-CLM and the linear regression approximation, averaged over all test cases from Latin Hypercube. The temporally averaged values (averaged for all months) are displayed for all quantities.

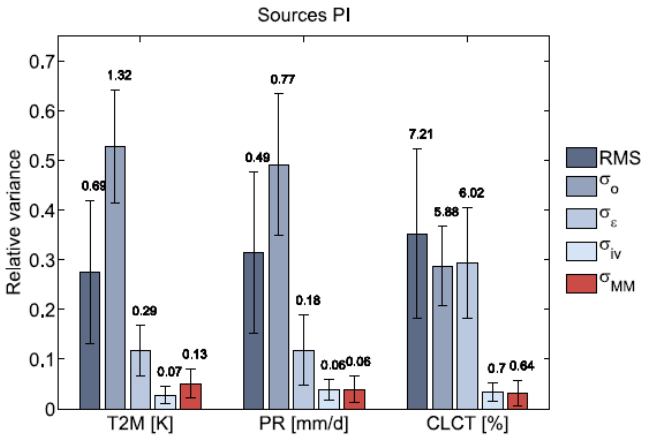

Figure 7. Comparison of the imprecision of the metamodel (σ_{MM}, red column) to all sources of the performance index (*PI*, blue columns) for T2M, PR and CLCT separately. The individual columns show the median values of all terms, whereas the error bars show the inter-quartile range derived from all spatial means considered to compute *PI*. To compare the terms for all three model variables all terms are scaled to sum up to 1 for each model variable. The original values are shown on top of each column with the dimension given by each model variable in the horizontal axis. The figure shows that the imprecision of the metamodel is small compared to the other sources of uncertainty in *PI*.

Figure 2: This plot was taken from Bellprat et al. 2012.

Taking these methodological distinctions into account, we draw careful conclusions. Temperature approximation errors are comparable in magnitude in both frameworks, ranging between 1.5 and 2.0 times the model's intrinsic variability. The quadratic regression, however, demonstrates notably higher precision for precipitation, achieving errors on par with intrinsic variability, versus approximately $1.6 \cdot \sigma_i$ for the linear Meta-Model. This indicates that the quadratic regression significantly improves precipitation approximation. Nevertheless, Fig. 1 clearly shows that the imprecision of the linear Meta-Model (green bars) is still much smaller than the typical error to observations (blue bars) for all variables except precipitation. This demonstrates the potential of optimization with linear emulator. The current manuscript shows that improvements in precipitation are close to the limit (see 'tune_prec' in Fig. 8 of the manuscript), since the optimized configuration yields only slightly lower RMSE for **pr_amount** than initial configuration. This may be due to imprecision in the Meta-Model, as well as the relatively low sensitivity of precipitation to the parameters considered (see the 'pr_amount' column in Fig. 5 of the manuscript).

The following paragraph is added to the section **Meta-Model validation** :

To assess the inaccuracy of the approximation statistically, we computed the monthly mean values of RMSE between the ICON-CLM output and the linear Meta-Model for each test case in the Latin Hypercube, and plotted the mean values in Fig. 1. As can be seen, the imprecision of the linear approximation (green bars) is slightly greater than the intrinsic uncertainty of the ICON-CLM (orange bars), by a factor of 1.5–1.7 for **tas**, **rsds**, **tasmin**, **tasmax** and **pr amount**, and by a factor of 2.5 for **hfls**. However, this imprecision (green bars) is still much smaller than the typical error to observations (blue bars) for all variables except precipitation, indicating the potential for optimization.

2. Another comment for the content is about L327 ("occasionally yielding negative precipitation values"). Doesn't this deserve more attention than a remark? We wonder whether this has consequences for the optimization, if you would advise something to go around this (e.g., add a conditional check in the regression to forbid out of range variable results and make those fall back to the range limit, in this case zero. I don't know, really, it's a proposition).

We thank the reviewer and agree that this point requires further clarification. While it would be possible to force the emulator to produce only non-negative precipitation, doing so would limit our ability to obtain the error-norm gradient analytically and complicate the optimization process. To maintain simplicity and efficiency, we have chosen not to add that constraint.

We originally assumed that negative precipitation had little effect and would like to prove it. Fig. 3 shows the regression and simulation output histograms. The regression consistently overestimates rainfall, which is the main source of inaccuracy. There are also a fair number of slightly negative values (within $[-10; 0]$ mm per month), but they contribute very little to the overall RMSE, as we will show next.

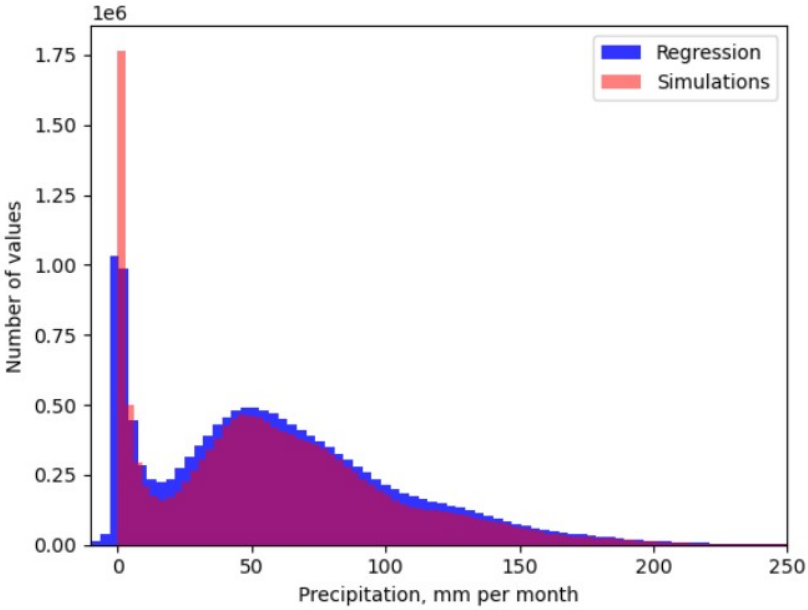

Figure 3: Histogram of precipitation values across all test setups sampled by Latin-Hypercube within the parameter bounds. This histogram was generated from the array of all monthly values at every grid point for all test configurations.

For each parameter set in our validation (sampled by Latin-Hypercube within the minimum and maximum bounds), we computed the $\mathrm{RMSE}\left(\mathrm{REG}^{\mathrm{pr}} - \mathrm{REG}^{\mathrm{pr}}_{\mathrm{pr}>0}\right)$, where $\mathrm{REG}^{\mathrm{pr}}$ is the unconstrained precipitation regression output and $\mathrm{REG}^{\mathrm{pr}}_{\mathrm{pr}>0}$ is the same output with all negative values replaced by zero. The average RMSE across all tests is approximately 0.2 mm per month, which is only $\sim 3\%$ of the intrinsic uncertainty in precipitation. Therefore, setting negative precipitation values to zero has a negligible effect on the LiMMo optimization.

The following paragraph is changed in the section **Meta-Model validation** from

A direct comparison between the regression model and the ICON-CLM simulation for different grid points and months is presented in Fig.6. For the variables **tas**, **tasmin**, **tasmax**, **psl**, and **hfls**, the coefficient of determination (R^2) exceeds 0.95 (not shown), indicating a decent approximation by the linear model. The variable **rsds** exhibits some dispersion around the mean but maintains a high determination coefficient. In contrast, precipitation (**pr_amount**) shows the poorest performance, with the Meta-Model occasionally yielding negative precipitation values, which compromises accuracy due to the lack of a constraint enforcing non-negative precipitation amounts.

to

A direct comparison between the regression model and the ICON-CLM simulation for different grid points and months is presented in Fig.6. Here values are plotted together for all test cases from Latin Hypercube. For the variables **tas**, **tasmin**, **tasmax**, **psl**, and **hfls**, the coefficient of determination (R^2) exceeds 0.95 (not shown), indicating a decent approximation by the linear model. The variable **rsds** exhibits some spread around the mean, but maintains a high determination coefficient(> 0.99). The precipitation (**pr_amount**) shows the poorest performance of all optimization variables. The spread exhibits values of up to 100 mm per month and the determination coefficient R^2 is 0.9 only. A comparison of the histograms (not shown) reveals that the Meta-Model yields slightly higher precipitation values than ICON-CLM. Also, due to the lack of physical constraints, the Meta-Model yields marginally negative precipitation values; however, their impact on the overall RMSE is very limited (approximately 3% of the intrinsic uncertainty of precipitation (Eq.2)).

3. The remaining of my comments are about the form. The most serious is about the introduction, which, in my opinion, is problematic.

We appreciate this comment and would like to improve the manuscript.

First, it relies on the abstract. For instance, the objective (designing a flexible cost-effective tuning tool) is not clearly stated (it is only the first sentence of the abstract), although all the decisions to reach it are developed from paragraph 2. In addition, the LiMMo acronym is employed without introduction (except from the abstract), relying on the implicit understanding from the reader about the paper's goal.

The introduction has been rewritten. The motivation behind improving the tuning framework is explained in the first paragraph, so our aim should be clear to the reader from the very beginning. The second paragraph provides a brief overview of existing tuning frameworks and introduces the LiMMo acronym. We hope the revised introduction is straightforward.

Second, the bibliography presented is very weak, emphasizing 2 examples of tuning techniques in the first paragraph, then 3 previous articles using quadratic regression, and the "well established" reference to justify the sample number formula in the context of Monte Carlo. There is no reference for Meta-Models in general, linear regression approaches and how it is perceived by previous Meta-Model studies (no mention in previous research would even be surprising enough to precisely mention that it's lacking), gradient-based optimizations, or the choice of objective functions ("RMSE is not enough" by Liemohn et al., 2021, for instance, is clearly relevant). You may then answer that some things are original ideas, but in this case, well there is a third point.

We have significantly expanded the list of literature. We provided a reference to the overview manuscript (Hourdin et al., 2017). The classification of existing methods is presented in the second paragraph. The third paragraph is devoted to discussing the objective calibration approach. We present the state of the art and emphasize the drawbacks that we aim to improve in the current manuscript. The linear regression approach and gradient-based optimization have rarely been applied to regional climate model tuning, as emphasized in paragraph 5. The choice of objective functions is briefly discussed in paragraph 4.

Third, many of the introduction is about explaining the choices made for LiMMo and stating about the advantages of the method. In my opinion, this is not introductory but rather about the methodology or even conclusion (for paragraphs L59 and L66 about the applicability of LiMMo).

The revised introduction now primarily discusses existing tuning techniques. The choices made for LiMMo are briefly mentioned as a way to overcome **objective calibration** problems.

In the end, the introduction just feels like a detailed abstract, which, I believe, is not what it should be. The introduction should present the field of your study (i.e., model parameter optimization), explain what has already been proposed in the literature (more extensively than in the current version, explain their procedure as you do for you: how do they choose the metrics, the weights, the optimization process, etc.), the pros, the cons. Then you explain the paper aims to fill the gap of flexible, cost-efficient parameter optimization by proposing a new method. You introduce BRIEFLY your choices, which you develop only in the methodology.

We agree that the initial version of the introduction did not cover the substantial literature. In the revised version, we provided a significantly extended list of literature. As previously described, we first provided the general motivation, followed by the literature overview. Next, we emphasized the current problems of objective calibration and suggested a new framework to address them. Finally, we briefly introduce the main features of LiMMo.

4. I'd actually suggest two different sections, one "materials" presenting ICON-CLM, its variables, tuning parameters and observational datasets, and one separate "the LiMMo framework" presenting the error norm, linear approximation, gradient-based optimization and the formulas of sensitivity. The separate section on LiMMo would facilitate the readability and applicability of your method by the readers, I believe. And then the "results" section is, in fact, entirely an application of your parameter optimization framework, from the sensitivity to the selection of parameters, to ICON-CLM.

We agree that separating the 'ICON-CLM description' and the 'LiMMo framework' would improve the readability of the paper. The old and the revised structures of the manuscript are presented in Fig. 4.

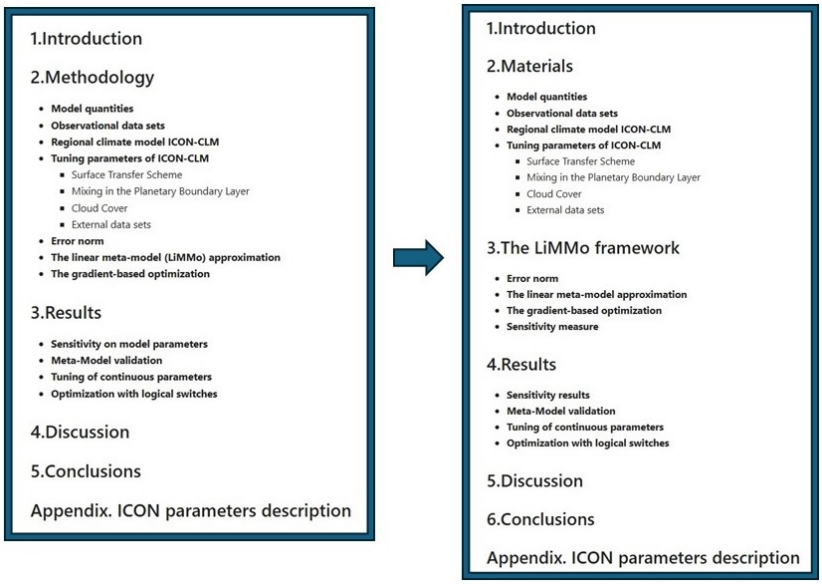

Figure 4: The initial (left) and revised (right) structure of the manuscript.

5. Now more specific comments:
L26-28: This relies on the implicit assumption that we understand what you are going to do, i.e., use Meta-Model. Then, in this context, you choose the regression-based approach. Please make it explicit, e.g., writing in the previous sentence: "This approach is referred as objective tuning, or objective calibration, and this is the focus of our study." In addition, the detail about "for each grid point and time step" seems fairly early in the text, and feels more like a methodology part explanation.

This part has been omitted from the revised introduction. The specific choices made for LiMMo are briefly presented in the introduction and explained in the section "**LiMMo Framework**".

Paragraph starting L29: The use of hyphens (which should be en or em dashes depending on the convention) for explaining the minimum number of required dynamic simulations is confusing in my opinion, because subjects are numbers and I'd very naturally read them "minus" at the first pass. Consider using colons, "i.e." or "that is", instead.

This part was omitted. We will only mention that the number of simulations for the quadratic approach scales as N^2. See paragraph 3 of the revised introduction.

L35,36: Your explanation corresponds to two formulas, i.e., two interpretations of the linear regression. If the reference simulation is fixed, then the formula is indeed N + 1, but if there is a new reference simulation each time you change of parameters, then the formula is simply 2 N. I think mentioning the 2 N option is confusing, especially since this is not the chosen approach.

This part was also omitted in the revised introduction. Nevertheless we would like to answer the comment. Indeed, if the reference configuration remains fixed, only $N + 1$ simulations are required. This is not fully the case for the current study, however. While the reference configuration changed over time, it did not change for every parameter. Nevertheless, $2N$ is more accurate for our simulations. Anyway, the main point is that the number of simulations is linear.

L51: Is it "perfect", really?

This part was also omitted in the revised introduction. Nevertheless we would like to answer the comment. As shown in Fig.3, the gradient of the error norm function approaches zero during the optimization process. This indicates that the local minimum of the error norm function has been achieved with very high precision (10^{-5} for the function increment by default). This is another advantage compared to the previously utilised Monte-Carlo approach, besides computational efficiency and linear scalability. In the case of a linear emulator with an RMSE error function, the error norm is actually a convex function. This implies that there is only one global minimum of the function, which is found perfectly by proposed gradient-based method.

Of course, the word 'perfect' might be slightly confusing, since the reader should also bear in mind the imprecision of the emulator itself. However, with a fixed linear emulator and an RMSE error norm, the proposed method could indeed be described as achieving 'perfect accuracy'.

L86: Consider using a table rather than a list

The list of model model quantities

- **tas**, hourly mean 2-meter temperature (deg K);
- **rsds**, hourly mean downward net short-wave radiation flux, ($W \cdot m^{-2}$);
- **tasmin**, daily minimum 2-meter temperature (deg K);
- **tasmax**, daily maximum 2-meter temperature (deg K);
- **psl**, hourly mean atmospheric pressure at the surface (Pa);
- **pr_amount**, hourly total amount of precipitations (mm per h);
- **hfls**, hourly mean surface downward latent heat flux ($W \cdot m^{-2}$).

is changed to the Tab.1

Table 1: The list of model quantities considered for tuning.

Acronym	Description	Unit
tas	hourly mean 2-meter temperature	K
rsds	hourly mean downward net short-wave radiation flux	$\mathrm{W} \cdot m^{-2}$
tasmin	daily minimum 2-meter temperature	K
tasmax	daily maximum 2-meter temperature	K
psl	hourly mean atmospheric pressure at the surface	Pa
pr_amount	hourly total amount of precipitations	mm per h
hfls	hourly mean surface downward latent heat flux	$\mathrm{W} \cdot \mathrm{m}^{-2}$

L98: Please highlight the nature of the E-OBS dataset (assimilating model, satellite-based product, satellite-station merge, ...)

The E-OBS dataset is a station-based, observational gridded dataset, not a satellite product nor a satellite-station merge. It is built from high-density in-situ measurements provided by over 2000 European meteorological and hydrological stations, interpolated onto a regular grid, and provided with ensemble uncertainty estimates.

The paragraph

As a reference for **tas**, **rsds**, **tasmin**, **tasmax**, **psl**, and **pr_amount**, the E-OBS version 29.0 data set (Cornes et al., 2018) was selected. It provides high quality daily data over Europe with a spatial resolution of about 25 km and a temporal coverage since 1950. With its fine spatial detail, daily temporal resolution, and ensemble-based uncertainty estimates, E-OBS is a robust resource for analyzing regional climate variability, long-term trends, and making reliable climate assessments.

is changed to

The E-OBS version 29.0 data set (Cornes et al., 2018) was selected as a reference for **tas**, **rsds**, **tasmin**, **tasmax**, **psl** and **pr_amount**. This land-only, station-based observational gridded data set is compiled from high-density in-situ measurements provided by over 2000 European meteorological and hydrological stations. These measurements are then interpolated onto a regular grid and provided with ensemble uncertainty estimates. It provides regularly updated high-quality daily data over Europe with a spatial resolution of approximately 25 km (12 km resolution is also available in the latest versions) and temporal coverage since 1950. Due to its fine spatial detail, daily temporal resolution and ensemble-based uncertainty estimates, E-OBS is a robust resource for analysing regional climate variability and long-term trends, and for making reliable climate assessments.

L105: Please provide the version and associated reference for the COARE algorithm.

Paragraph

We aim to tune the **hfls** to align with the HOAPS version 4.0 data set (Andersson et al., 2010). HOAPS provides a satellite-based climatology of latent heat flux over the global ice-free oceans, derived from recalibrated SSM/I and SSMIS sensor measurements. It covers the period from 1987 to 2014 with a spatial resolution of about 55 km and provides 6-hourly averages. Using the COARE bulk flux algorithm, HOAPS provides accurate estimates, making it a key reference for ocean-atmosphere interaction studies and energy exchange assessments.

is changed to

Our aim is to calibrate the **hfls** to align with the HOAPS version 4.0 data set (Andersson et al., 2010). HOAPS provides a satellite-based climatology of latent heat flux over the global ice-free oceans, derived from recalibrated SSM/I and SSMIS sensor measurements. The data set covers the period from 1987 to 2014, has a spatial resolution of approximately 55 km, and provides 6-hourly averages. HOAPS uses the COARE bulk flux algorithm version 2.6a (Fairall et al., 2003), to provide accurate estimates, making it a key reference for ocean-atmosphere interaction studies and energy exchange assessments.

L219: Please mention somewhere (for instance in the introduction) existing objective ways to define weights in Multi-Criteria Decision-Making (e.g., entropy weights)

The question of different weighting strategies is briefly discussed in the **Discussion** section (paragraph from L385) of the manuscript. Indeed, there is a great deal of flexibility in how the weights are defined. This is a very interesting and important topic that requires careful investigation. As can be seen in Fig. 1, the signal-to-noise ratio differs for different model quantities. The temperature quantities (**tas**, **tasmin** and **tasmax**) and latent heat flux over the sea (**hfls**) exhibit similar values of 12–18, while radiation (**rsds**) is slightly smaller and precipitation (**pr_amount**) is significantly smaller. This may be because the NWP configuration is already very well tuned for precipitation. We experimented by assigning weights that were inversely proportional to the signal-to-noise ratio of the initial setup (similar to entropy weights in information theory), but we decided not to draw any firm conclusions from the results and to leave this area of research for future investigation. In any case, we have demonstrated that the result depends heavily on the definition of the weights. Therefore, the 'objective tuning' approach is actually very subjective and depends heavily on the user's objectives. Indeed, some

objective strategies for Multi-Criteria Decision-Making like entropy weights might be applicable, but this is still open discussion.

Another aspect is the weights of monthly values in the error norm definition. In the case of 12-km spatial resolution of climate model, the monthly noise is fairly homogenous for all seasons, therefore simply taking the monthly average in error norm equally treats all seasons. This is not the case for finer grid (e.g., convection permitting simulations with 3-km spatial resolution). We are currently preparing another manuscript with LiMMo optimization for this case, but leave all this questions out of current work. The main idea of current manuscript is to present the new tool, describe its functionality and conduct very preliminary experiments with weights.

Anyway, we agree that the entropy weights worth to be mentioned. Therefore, following paragraph is added to the section **Tuning of continuous parameters**

There are also some objective ways of defining weights, such as entropy weights for multi-criteria decision-making in information theory, which are beyond the scope of the current study. These could be implemented in the LiMMo framework by assigning a variable weight that is inversely proportional to signal-to-noise values of the initial configuration for each model quantity.

L255: Please split the paragraph after single norm evaluation O(...) and an additional sentence clarifying that the optimization then seeks for the vector p that minimizes Eq. 11. Address the optimization method in another paragraph.

The paragraph

The computation of the gradient requires one loop over grid points (i, j), time (k), and model variables (n), making its duration comparable to that of a single norm evaluation $O(N_x \cdot N_y \cdot N_t \cdot N_{\text{vars}})$. The availability of a fast gradient computation procedure allows the use of different optimization methods. This study proposes the implementation of the Limited-memory Broyden-Fletcher-Goldfarb-Shanno with Box constraints (L-BFGS-B) algorithm (Broyden, 1970; Byrd et al., 1995). This method is chosen due to its high convergence speed, being a quasi-Newton method that approximates the Hessian matrix, and its capability to impose constraints on parameter ranges, thereby eliminating nonphysical parameter values during the optimization.

is changed to

The computation of the gradient requires one loop over grid points (i, j), time (k), and model variables (n), making its duration comparable to that of a single norm evaluation $O(N_x \cdot N_y \cdot N_t \cdot N_{\text{vars}})$.

The availability of a fast gradient computation procedure allows the use of different optimization methods. Gradient-descent-type optimization involves iterations over the vector of parameters **p** that search for the minimum error norm function (Eq. 4) in the direction opposite the gradient (Eq. 11).

This study proposes the implementation of the Limited-memory Broyden-Fletcher-Goldfarb-Shanno with Box constraints (L-BFGS-B) algorithm (Broyden, 1970; Byrd et al., 1995). This method is chosen due to its high convergence speed, being a quasi-Newton method that approximates the Hessian matrix, and its capability to impose constraints on parameter ranges, thereby eliminating nonphysical parameter values during the optimization.

Relating to the axes and titles in Fig. 3 and 4, please specify clearly that "Score = Eq. 4 = Objective function value", and that "Score gradient = Gradient norm value = Eq. 11" and that iterations are made with the p vector (or something else, if I misunderstood). Otherwise, please make the terms more consistent.

We agree, that the terms might be confusing, therefore the Y-axis labels are changed in Fig.3 and Fig.4 from "Objective function value" to "Error norm" and from "Gradient norm value" to "l2-norm of gradient". The title of Fig.3 and Fig.4 is changed from "Score" to "Error norm" and from "Score gradient" to "Error norm gradient".

The caption of Fig.3 is changed from

Figure 3. Convergence of L-BFGS-B method: score function values without (a) and with (c) parameter normalization, l2-norm of score gradient without (b) and with (d) parameter normalization.

to

Figure 3. Convergence of L-BFGS-B method: error norm (Eq.4) values without (a) and with (c) parameter normalization, l2-norm of error norm gradient gradient (Eq.11) without (b) and with (d) parameter normalization.

The caption of Fig.4 is changed from

Figure 4. Ensemble of 15 optimization trajectories: (a) score function, (b) l2-norm of score gradient.

to

Figure 4. The axes are the same as in Fig. 3. Ensemble of 15 optimization trajectories with disturbed initial conditions: (a) error norm (Eq.4), (b) l2-norm of error norm gradient (Eq.11).

Eq. 14: I'd highly prefer to see the sensitivity benchmark in the methodology section. In the current version, those case-independent equations feels odd after having introduced a result section.

The formulas for sensitivity measure are moved to the section **3.The LiMMo framework**, subsection **3.4 Sensitivity measure**.

Fig. 5: Please use a uniform decreasing intensity colormap such as cmocean's amp here. The diverging one of the current version makes no sense because the center (white) is not indicated and has no meaning anyways. Moreover, please address the contrast between text and background (use white text color under a certain threshold of background intensity). Also, consider using the average rather than the sum, so as to include the column in the coloring.

This comment is indeed very helpful in improving the readability of the sensitivity table. Please refer to the 'old' sensitivity table (Fig. 5) and the 'updated' sensitivity table (Fig. 6).

	tas	rsds	tasmin	tasmax	psl	pr_amount	hfls	Sum
taw1	7.235	2.940	5.190	6.683	3.473	1.838	2.235	29.594
taw2	0.730	0.842	0.742	0.739	0.784	0.940	0.937	5.713
rlh	2.047	1.447	2.173	2.060	2.410	1.711	5.067	16.914
rs	2.561	2.218	2.778	2.270	3.061	2.767	6.156	21.813
rl	1.761	2.577	1.837	1.809	1.749	2.132	2.145	14.010
rsmf	2.163	2.268	2.069	2.247	2.032	2.441	2.530	15.751
crb	1.668	1.328	1.576	1.682	1.269	1.445	1.698	10.667
tbl	2.694	6.281	2.733	3.019	3.143	2.595	3.018	23.482
tbla	2.329	5.797	2.100	2.811	2.835	1.733	2.318	19.924
ao	3.826	7.540	2.723	4.281	3.524	2.266	3.232	27.393
aot4	2.179	3.241	1.761	2.274	1.590	1.590	2.016	14.652
tkhmin	2.571	1.203	3.416	1.760	1.366	1.306	1.342	12.964
sdb	2.358	0.833	4.093	1.434	0.925	0.984	1.666	12.292
acrf	0.954	1.249	0.917	0.973	0.940	0.997	0.942	6.971
oro	1.794	0.903	1.968	1.448	1.272	1.110	1.455	9.950

Figure 5: The sensitivity measure of prognostic variables (columns) on model parameters (rows) computed as Eq.15. The last column gives the sum in the row, which shows the overall sensitivity of the model to the parameter. The numbers are colored in a "blue to red" palette with increasing values.

	tas	rsds	tasmin	tasmax	psl	pr_amount	hfls	Avg
taw1	7.235	2.940	5.190	6.683	3.473	1.838	2.235	4.228
taw2	0.730	0.842	0.742	0.739	0.784	0.940	0.937	0.816
rlh	2.047	1.447	2.173	2.060	2.410	1.711	5.067	2.416
rs	2.561	2.218	2.778	2.270	3.061	2.767	6.156	3.116
rl	1.761	2.577	1.837	1.809	1.749	2.132	2.145	2.001
rsmf	2.163	2.268	2.069	2.247	2.032	2.441	2.530	2.250
crb	1.668	1.328	1.576	1.682	1.269	1.445	1.698	1.524
tbl	2.694	6.281	2.733	3.019	3.143	2.595	3.018	3.355
tbla	2.329	5.797	2.100	2.811	2.835	1.733	2.318	2.846
ao	3.826	7.540	2.723	4.281	3.524	2.266	3.232	3.913
aot4	2.179	3.241	1.761	2.274	1.590	1.590	2.016	2.093
tkhmin	2.571	1.203	3.416	1.760	1.366	1.306	1.342	1.852
sdb	2.358	0.833	4.093	1.434	0.925	0.984	1.666	1.756
acrf	0.954	1.249	0.917	0.973	0.940	0.997	0.942	0.996
oro	1.794	0.903	1.968	1.448	1.272	1.110	1.455	1.421

Figure 6: The sensitivity measure of prognostic variables (columns) on model parameters (rows) computed as Eq.15. The 'Avg' column shows the mean sensitivity of the model to the parameter, calculated as the mean values in the rows. Darker shades are used to color the background of the numbers for larger values.

L301: "The heat flux..." I find that there are too many exceptions in the figure (rsds is the lowest for rlh, or pr is quite high as well although not mentioned) to state this that way.

We agree with this comment. The paragraph

Overall, the sensitivity results are consistent with theoretical expectations. It is clear that the surface albedo parameterization (**taw1**) is the primary driver of surface air temperature variations (**tas**, **tasmin**, **tasmax**). The heat flux scaling factors (**rlh**, **rs**, **rl**) show sensitivity to both short-wave radiation (**rsds**) and latent heat flux over the sea (**hfls**) with considerable impact on temperature quanities (**tas**, **tasmin** and **tasmax**). The soil resistance parameters (**rsmf** and **crb**) exhibit sensitivity across all model variables. Although optimizing these parameters may not lead to improvements in one variable without affecting others, their inclusion may still be beneficial for optimization.

is changed to

Overall, the sensitivity results are consistent with theoretical expectations. It is clear that the surface albedo parameterisation **taw1** is the primary driver of surface air temperature variations (**tas**, **tasmin**, **tasmax**). **taw2** has a negligible impact on the model variables, which is below the level of the ICON-CLM's intrinsic variability. The heat flux scaling factors **rlh** and **rs** show sensitivity primarily to latent heat flux over the sea (**hfls**), with a moderate impact on other quantities. The ratio of the laminar scaling factors **rl** has the greatest impact on short-wave radiation (**rsds**), contributing only slightly to precipitation (**pr_amount**) and latent heat flux (**hfls**). The soil resistance parameters **rsmf** and **crb** exhibit sensitivity across all model variables. Although optimizing these parameters may not lead to improvements in one variable without affecting others, their inclusion may still be beneficial for optimization.

L320: You mean one single sample vector p, right? (it's a single simulation in Fig. 6, correct?)

We selected the following parameters to assess the quality of linear approximation: **taw1**, **rlh**, **rs**, **rl**, **tbl**, **tbla**, **ao** and **tkhmin**. These form the vector of parameters **p**. We then sampled seven different vectors **p** from Latin-Hypercube between the minimum and maximum values for each parameter, and simulated these seven configurations. Unfortunately, we do not have the computational resources to extend the validation samples. Each dot in Fig.6 shows the monthly value of a single grid point for a specific configuration. Ultimately, all monthly values, grid points, for all simulated validation configurations are displayed.

We added the following sentence to paragraph L323

A direct comparison between the regression model and the ICON-CLM simulation for different grid points and months is presented in Fig. 6. Here values are plotted together for all test cases from Latin Hypercube.

L321: "around tthe minimum and maximum values" What does this mean? The parameters were not taken out of their range, were they?

This is a typo, thank you for noticing it. The tested values were taken within the min/max range. The sentence

Test samples were generated by simultaneously varying these parameters within the Latin Hypercube around the minimum and maximum values.

is changed to

Test samples were generated by simultaneously varying these parameters from the Latin Hypercube within the minimum and maximum values.

L321: "Due to limited..." It is unclear whether this subset is the same as in this paragraph's second sentence or a additional filtering within this subset.

Indeed, it is not clear from the text. This was an explanation why only parameters **taw1**, **rlh**, **rs**, **rl**, **tbl**, **tbla**, **ao** and **tkhmin** were selected (not all parameters presented in **Materials** section). We rephrased the paragraph

To evaluate the accuracy of the linear Meta-Model approximation, several parameter configurations were simulated with ICON-CLM. Some of the most influential parameters identified from the sensitivity analysis (Fig. 5) were selected: **taw1**, **rlh**, **rs**, **rl**, **tbl**, **tbla**, **ao** and **tkhmin**. Test samples were generated by simultaneously varying these parameters within the Latin Hypercube around the minimum and maximum values (see Tab. A1 and Tab. A2). Due to limited computational resources, only the subset of the most sensitive parameters was considered.

as

Several parameter configurations were additionally simulated with ICON-CLM to evaluate the accuracy of the linear Meta-Model approximation. Due to limited computational resources, only a subset of parameters was considered. The most influential parameters, which exhibited the largest sensitivity in the sensitivity analysis (see Fig. 5), were selected: **taw1**, **rlh**, **rs**, **rl**, **tbl**, **tbla**, **ao** and **tkhmin**. Test samples were generated by simultaneously varying these parameters from the Latin Hypercube within the intervals from minimum to maximum values (see Tab. A1 and Tab. A2).

Fig. 7: I do not understand what are the markers for. Is that for several sets of parameters? Please clarify this because it is not easily understandable after Fig. 6. Also, consider bigger markers or other shapes. It is currently difficult to distinguish between round-like markers (penta and hectagonal).

Here, each marker represents a separate validation setup from Latin Hypercube (i.e., one of the seven tested values of the parameter vector **p**). We increased the size of the markers in Fig.7. We also made the explanation clearer by rephrasing the paragraph

The linear approximation error for various variables was assessed by comparing the time-averaged (averaged over all climatological months) RMSEs with the observations (Eq.1), as shown in Fig.7. The scores of the dynamical simulations and their corresponding Meta-Model approximations are represented by markers of identical shape. Notably, the distance between almost all pairs of markers with the same shape across the axes remains within the range of the intrinsic variability (Eq.2) of the climate model. With a few exceptions, the order of the linear and dynamic errors is largely maintained. This indicates that the optimal linear approximation closely matches the optimal ICON-CLM configuration with a high degree of accuracy, especially when the RMSE is reduced by an amount exceeding the intrinsic variability of the variable under consideration.

as

The linear approximation error for various variables was assessed by comparing the time-averaged (averaged over all climatological months) RMSEs with the observations (Eq.1), as shown in Fig.7. For each of the Latin Hypercube validation setups, we plot the RMSE to observations for different pairs of variables, for both the linear regression approximation and the corresponding dynamical simulation. The scores of the dynamical simulations and their corresponding Meta-Model approximations are represented by markers of identical shape. With a few exceptions, the order of the RMSEs for the linear and dynamic models is largely maintained, i.e., if the RMSE is smaller for the regression results, the same is true for the dynamical simulation. This justifies the reduction in the RMSE-based error norm for the linear emulator, which is minimised by the optimisation procedure, corresponding to an improved dynamic setup with reduced biases. This is particularly true when the reduction in RMSE exceeds the level of imprecision in the approximation, bearing in mind the error in the linear approximation.

Section 3.4: Please remind the readers that introducing logical switches does not affect the optimization results for continuous parameters, and that the process simply consists of computing the error using the regression Eq. 9 for all new eight possibilities.

In fact, the addition of logical switches affects the optimization results. Adding a logical switch to Eq. 9 introduces a constant shift in the Meta-Model. While these switches do not impact the gradient Eq. 11, the shifted function is minimized during the optimization procedure. This results in a different optimal vector of parameters **p**.

We clarified this aspect by changing paragraph

This subsection presents the optimization results obtained using the Meta-Model with incorporated logical switches (Eq.9). The parameter set is fixed as in the previous subsection, with the '**expert_weights**' weight configuration applied. The study considers three logical parameters (**sdb**, **acrf** and **oro**), resulting in a total of eight possible configurations. For each configuration, the continuous parameters were optimized. The results are summarized in Fig.9. The final scores table provides the comprehensive information needed to make an objective decision in selecting the climate model configuration that best meets the user's priorities and interests.

to

This subsection presents the optimization results obtained using the Meta-Model with incorporated logical switches (Eq.9). The parameter set is fixed as in the previous subsection, with the '**expert_weights**' weight configuration applied. The study considers three logical parameters (**sdb**, **acrf** and **oro**), resulting in a total of eight possible optimizations. The continuous parameters were optimized for each configuration of logical switches that defines the shifted linear Meta-Model. The results are summarized in Fig.9. This final scores table provides the comprehensive information needed to select the climate model configuration that best meets the user's priorities and interests.

L414: "objectively" does not make sense if it's adapted to the user's priorities.

True, the word "objectively" is excluded from the sentence.

References

Bellprat, O. et al. (2012). "Objective calibration of regional climate models". In: *Journal of Geophysical Research: Atmospheres* 117.D23. DOI: 10.1029/2012JD018262. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2012JD018262. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012JD018262.

Rockel, B., A. Will, and A. Hense (2008). "The Regional Climate Model COSMO-CLM (CCLM)". In: *Meteorologische Zeitschrift* 17.4, pp. 347–348. DOI: 10.1127/0941-2948/2008/0309.