<u>Point-by-point response to referee comments</u>

Surrogate-based model parameter optimization in simulations of the West African monsoon

<u>Matthias Fischer</u>, Peter Knippertz, Carsten Proppe

We would like to thank the reviewers for their constructive and helpful comments on the manuscript. Overall, we agree with the given remarks and provide a short response below.

RC1:

The main issue in my view is that there have been various similar proof-of-concept studies of objective calibration of weather, climate, ocean or cryospheric models in the past. I don't see any major conceptual or methodological innovation in the present work, except maybe the use of Principal Component Analysis in the dimension reduction process, although Gaussian process models have certainly been used extensively (An internet search suggests that PCA has been used previously, too, such as in: Chang, W., Applegate, P. J., Haran, M., and Keller, K.: Probabilistic calibration of a Greenland Ice Sheet model using spatially resolved synthetic observations: toward projections of ice mass loss with uncertainties, Geosci. Model Dev., 7, 1933–1943, https://doi.org/10.5194/gmd-7-1933-2014, 2014). It is good to be reminded of advantages and drawbacks of objective model calibration in the context of weather and climate modelling, but similar studies can be performed using a multitude of models, regions, parameters, and loss functions; there should be a particular added value in such an effort, either in terms of methodology or process understanding.

There is a long history of using surrogate models (more often called "emulators" in the weather and climate community) to calibrate weather and climate models, the relevant literature is not quite adequately reviewed in the present manuscript.

There are also various software packages in this context, e.g. Watson-Parris, D., Williams, A., Deaconu, L., and Stier, P.: Model calibration using ESEm v1.1.0 – an open, scalable Earth system emulator, Geosci. Model Dev., 14, 7659–7672, https://doi.org/10.5194/gmd-14-7659-2021, 2021; Couvreux, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque, N., et al. (2021). Process-based climate model development harnessing machine learning: I. A calibration tool for parameterization improvement. Journal of Advances in Modeling Earth Systems, 13, e2020MS002217. https://doi.org/10.1029/2020MS002217.

Nowadays studies often use machine learning methods in the process of building emulators, which can be viewed an innovation to some extent.

I like the honesty of the conclusions in the present manuscript, but there are various past proof-of-concept studies with very similar conclusions and I struggle to see any substantial innovation in the present work. Whether one can say that the incorporation of the approximation of 2D fields with Principal Component Analysis (PCA) in the optimization process is an innovation I'm not quite sure, certainly Gaussian process models have been used extensively (see e.g. Watson-Parris et al., 2021; Couvreux et al. 2021). PCA can be seen as a Gaussian process model, too, in a sense. In machine learning models dimension reduction is often performed using autoencoders, something I would consider more of an innovation in the model calibration context.

AC:

We appreciate the reviewer's comments and fully agree that the use of surrogate models in parameter tuning is not novel. Many previous studies have indeed explored the opportunities and limitations of such methods. However, we aim to advance these methodologies in specific and meaningful ways:

- *Multi-Objective Studies: It is common for multiple objectives to require optimization simultaneously, and this often necessitates a robust framework for handling these competing goals. While some previous works have employed fixed weights, our study emphasizes the importance of a more nuanced approach that accounts for both weight uncertainties and systematic weight variations. We believe this provides modelers with deeper insights into the relationships and trade-offs between objectives, an area that we found to be less explored in the meteorological literature.*

- *Integration of Meteorological Phenomena: To our knowledge, there is no existing literature that integrates the West African Monsoon (WAM) with the ICON model in the same detailed manner as our study. Since the ICON model has not been specifically optimized for the African continent, our results offer valuable insights into both the model's current capabilities and potential areas for improvement in relation to the WAM system.*

- *Application of Principal Component Regression: In the context of parameter tuning, we demonstrate that principal component regression is a promising tool. The interactive tool developed in this study provides a computationally efficient and insightful option for model developers, which can be applied in other contexts as well.*

We also acknowledge the reviewer's point regarding the relatively limited scope of our literature review on surrogate models for parameter tuning. We will expand this section by incorporating additional relevant and useful literature to provide a more comprehensive overview. While we do not claim to have reinvented surrogate-based parameter tuning, we believe our study introduces innovative elements that extend the conventional methodologies in meaningful ways.

regions or seasons while degrading it in others, or may enhance certain forecast variables at the expense of others. Traditionally, parameter tuning has been conducted by experts without a unified framework. History matching (Williamson et al., 2013) has emerged as a prominent technique for quantifying parameter uncertainties by systematically exploring a range of plausible model configurations and eliminating parameter sets that fail to reproduce observations within acceptable tolerances. This approach is particularly useful in the tuning of climate model parameters, such as those governing aerosol-cloud interactions, where direct measurements are often challenging to obtain (e.g., Lee et al., 2016). Over the past decades, automatic calibration techniques have emerged, incorporating data assimilation methods into operational weather forecasts. A comprehensive review by ~~Ruiz et al. (2013)~~ Ruiz et al. (2013) focused on these techniques and ~~Zängl (2023)~~ Zängl (2023) described the current implementation in the ICON model. ~~Ruiz et al. (2013)~~ Ruiz et al. (2013) emphasized that objective optimization often becomes infeasible when complex numerical models and a large number of parameters are involved. To address this, surrogate-based optimization techniques have gained attention in meteorological studies. These methods replace the expensive numerical models with cheaper surrogates, ~~also known~~ often referred to as meta-models or emulators, which can be trained with fewer NWP evaluations. For ~~instance, Neelin et al. (2010) and Bellprat et al. (2012) used a quadratic meta-model~~ example, quadratic meta-models were used to approximate climatic variables ~~.~~ by Neelin et al. (2010) and Bellprat et al. (2012). Similarly, Ray et al. (2015) employed Gaussian process and polynomial regression for Bayesian calibration of the Community Land Model. Chang et al. (2014) utilized principal component analysis in conjunction with Gaussian process regression to develop surrogate models for calibrating the Greenland Ice Sheet model. In another study, Lu et al. (2018) applied advanced sparse grid interpolation as a surrogate model for the E3SM Land Model, using quantum-behaved particle swarm optimization to identify optimal parameters. Various software packages have also been developed to facilitate specific optimization procedures,

such as the toolkit by Watson-Parris et al. (2021), which supports model calibration using a range of surrogate models (e.g., Gaussian process regression), and the parameterization improvement tool by Couvreux et al. (2021), which incorporates Gaussian process regression, history matching, and other techniques.

[...]

for all other atmospheric variables. ~~Through our novel approach, we determine~~ Parameter optimization in meteorological modeling has often not focused on the simultaneous consideration of multiple model outputs, and no general framework has been established for this purpose. In this study, we present a novel approach for determining optimal model parameters by incorporating ~~uncertainty in predefined weights for the QoIs~~ uncertainties in the predefined weights of quantities of interest (QoIs) and output fields, and by ~~examining the effect~~ systematically examining the effects of varying these weights. ~~This~~ To the authors' knowledge, no previous meteorological studies have explored the impact of weight variations in MOO in this manner. Also, no studies have conducted such an extensive exploration of model parameter calibration specifically for the West African monsoon system. Our approach, integrating both statistical and expert-driven perspectives, offers a promising pathway to improve simulations of the WAM system and to broaden our understanding of it.

~~To the best of the authors' knowledge, no meteorological studies have investigated variations in weights of objectives in MOO in such a manner. Additionally, in the context of surrogate-based optimization of meteorological variables, no studies have incorporated the approximation of 2D fields with PCA in the optimization process.~~

[...]

~~All optimization problems were solved~~ Several software packages for parameter calibration have been developed in recent years (e.g., Watson-Parris et al., 2021; Couvreux et al., 2021), and these tools are generally applicable across various contexts. However, given the specific requirements of this study—such as input space transformations, the use of ansatz functions for universal kriging, and variations in objective function weights—this work employs a combination of widely adopted Python packages, which have been adapted to meet the study's goals. Optimization tasks were performed using the *SciPy* package for Python~~and~~ ~~.~~, specifically utilizing the Nelder-Mead method (Virtanen et al., 2020).

RC1:

There are issues in particular with calibrating physical parameterisations in regional model simulations: it is usually not very difficult to improve physical parameterisations when studying a particular region (considering specific objective functions). The main difficulty in developing phyical parameterisations is that they need to be valid across a wide range of climates and meteorological conditions, from high to low latitudes, from wet to dry conditions etc. This is particularly obvious in the discussion of the convective entrainment rate in the present manuscript. Entrainment is a key process in convection and has been studied extensively. It is well known that in weather and climate models entrainment rates in convection parameterisations can affect weather and climate in various ways. E.g. Zhu, H. and Hendon, H.H. (2015), Role of large-scale moisture advection for simulation of the MJO with increased entrainment. Q.J.R. Meteorol. Soc., 141: 2127-2136. https://doi.org/10.1002/qj.2510; Sherwood, S. C., D. Hernández-Deckers, M. Colin, and F. Robinson, 2013: Slippery Thermals and the Cumulus Entrainment Paradox. J. Atmos. Sci., 70, 2426–2442, https://doi.org/10.1175/JAS-D-12-0220.1.

AC:

We appreciate the comment regarding the challenges of parameter tuning, especially for the entrainment rate. While we have already addressed the limitations and potential of parameter optimization in our study, we have now included additional details specifically addressing the entrainment parameter to provide further clarity.

also resulted in a localized precipitation increase along the central axis of the rain belt, suggesting a concentration of rainfall in areas with favorable ambient conditions. The challenge of tuning the entrainment rate, particularly in the tropics, has often been highlighted. For instance, low entrainment rates have been associated with higher climate sensitivities due to enhanced deep convection and increased moisture transport into the upper troposphere, potentially leading to an unrealistic cloud effect (Stainforth et al., 2005; Sanderson et al., 2008; Sexton et al., 2011). Additionally, Zhu and Hendon (2015) showed that an increased entrainment can improve Madden-Julian Oscillation simulations by modulating convection, although the model may still fail to capture the essential moistening by shallow convection. Changes in *zvz0i* resulted in similar effects but with some differences in the spatial distributions and magnitudes, particularly regarding high- and mid-level cloud cover. The param-

[…]

reducing entrainment rates pushing the optimal values almost to the lowest plausible level. Given that we suspect an influence of topographic rainfall enhancement (see Sect. 3.2), this may lead to a better agreement but not necessarily a physically more realistic model configuration. ~~In contrast to that, the~~ This underscores the challenge of tuning the entrainment rate in tropical regions. The other two parameters, *zvz0i* and *c_soil*, tend to converge towards values close to the means of their original PDFs,

RC2:

A first comment is that, even if the work is well described, probably suffers from the fact that some contents are based also on the results of research previously published by the same authors, and a lack of smooth reading could be envisaged. This is not necessarily an issue, since in this way it is preserved the briefness of the work, but in some points some recalls could helps readers.

[…]

the space (reduced) of 3 parameters is still well represented by the sampling of a 60 samples in a space of 6 variables? Since the training strategies is crucial in SBO and PCA technique, some clarification should be given in the current paper. […] maybe less than 60 training runs could be sufficient. With the remaining 'computational budget', an attempt in updating Surrogate models could be verified, adopting an approach that mitigate the risk of low reliability at first stages, that could affect also the optimal set of parameters […] have the authors verified that the set of optimal parameters found effectively give rise to better ICON results? for example by running icon with optimal set found.

AC:

We acknowledge that certain sections of the manuscript build upon our previously published research, which may have affected the overall flow for readers unfamiliar with the prior work. To improve clarity, we will ensure that key elements from the original paper are briefly recalled at relevant points, enhancing the reader's understanding while maintaining the manuscript's conciseness.

In our original work, we conducted computationally intensive ICON model simulations, which required several hundred thousand CPU hours for all years and parameter configurations. Given these high computational costs and the already limited experimental design (60 training points for 6 input dimensions), it was not feasible to exclude certain parameters via “cheap” simulations beforehand. Consequently, we employed a space-filling design with all 6 parameters.

Based on the results from sensitivity and parameter studies, it is indeed apparent that excluding the three parameters with minimal effect and reallocating the computational budget (or even less) to create a more accurate surrogate model for the remaining three parameters would be a promising approach. This could potentially involve the use of sequential training algorithms. However, the computational budget has already been expended for the original studies and the identification of parameter effects. Any alternative strategy would thus entail significant additional computational effort. Therefore, our focus remains on utilizing the models we have already developed without requiring further model simulations.

In our original work, we validated the Gaussian Process Regression models and demonstrated very good model accuracy. Although accuracy may vary throughout the input domain, the accuracy across the full 6-dimensional input space also covers the 3-dimensional subspace used in the optimization process. Hence, we can rely on the original validation results and the surrogate models employed.

As discussed in our manuscript, the overall improvement of the ICON model is relatively constrained, and depends on the weights of the objectives. With the high accuracy of the surrogate models based on validation results, Figure 5 illustrates possible changes in relation to the applied weights and their associated uncertainties.

In the manuscript, we will provide additional details on the step where the 6-dimensional surrogate models are applied to the 3-dimensional subspace, including further explanations of the reasons behind this approach and its limitations.

QoIs computed from these fields, which represent key characteristics of the WAM, are listed in Table 1. ~~Surrogate models were developed by defining~~ The experimental design consists of 60 training points $\mathbf{x}_i(i = 1 \ldots 60)$, i. e. ~~,~~combinations of the ~~model parameters, and conducting ICON simulations for all training points and all four years. The results~~ six considered ICON model parameters. These parameters were chosen based on expert judgment, as they were expected to have a substantial impact on WAM quantities. For each training point i, the QoIs $y_{ij}(i = 1 \ldots 60)$ ~~from the computation of QoI j were used to train the surrogate model which describes~~ were computed based on the temporally averaged 2D output fields. Surrogate models were developed using Gaussian process regression to describe a relationship between the six model parameters and each QoI j. ~~More detail~~ Our previous study demonstrated that only the three parameters entrainment rate (*entrorg*), terminal fall speed

of ice crystals (*zvz0i*) and soil moisture evaporation fraction (*c_soil*) had a substantial impact on the monsoon system. More details on the model setup ~~is~~ and uncertainty contributions are given in Fischer et al. (2024). Consequently, the optimization process in this study is limited to variations in these three parameters to minimize the risk of overfitting, as including parameters with low sensitivity would not produce meaningful adjustments. In hindsight, building surrogate models based solely on an experimental design that perturbed these three parameters would have been advantageous. That approach would likely yield higher model accuracy by concentrating the same number of training points in a lower-dimensional input space. However, since the simulation results had already been obtained in our prior study at considerable computational expense, we opted to use the existing data for the optimization studies to avoid further computational costs. The surrogate models are validated within the six-dimensional input space and are thus considered valid within the three-dimensional subspace for this study. Parameters not included in the optimization process are kept at their default ICON model values.

RC2:

Cm in formula 7 and 8. To my knowledge, that scalar terms are related to the inner product of PCA modes with the scalar fields, and the former are sorted according to their energetic content. Here the authors are trying to reduce the truncation error: maybe this requires more discussion and argumentation also in the text. […]

Trying to explain if this approach have could have an impact on the physical meaning of that scalar coefficient, e.g, considering that the energy content of the first 3 modes is related to the energy content expressed by the 3 first eigenvalues: are the new resulting scalar coefficients modifying this, since it can be demonstrated that the PCA scalar vector is linked to the relative eigenvalue, and hence to the energy content.

AC:

The coefficients Cm are obtained by solving a general minimization problem aimed at reducing the overall prediction error. This approach, commonly employed in regression models, ensures that the coefficients are optimized to fit the model to the data, even though they may not directly correspond to PCA eigenvalues or be associated with energy content. Nevertheless, the principal components are ordered in descending sequence based on their associated eigenvalues, with each eigenvalue reflecting the amount of variance explained by the corresponding component. The input space transformation T_ros is used to improve model performance by addressing distortions in the input space. While PCA coefficients can be analytically computed when using a linear ansatz for the coefficients, in our case, solving the minimization problem numerically was necessary and computationally feasible. As discussed in our original paper, the input space transformation enhances the surrogate modeling process. Although it may affect the traditional interpretation of the coefficients with respect to energy content, it does not compromise the validity or efficacy of the approach.

In the manuscript, we will include additional information to provide a clearer understanding of our methodology in the broader context of PCA. Furthermore, we will expand on the bibliographical background and offer more detailed explanations regarding the methodological aspects.

optimization strategy, surrogate models for the full 2D output fields are determined and used in the optimization process. Principal component regression is employed as a surrogate modeling technique for approximating 2D fields due to its effectiveness and interpretability. In contrast, neural network-based approaches, such as autoencoders, typically require substantially larger datasets, which are not feasible in this case due to the high computational cost of running ICON model simulations to generate training data.

Traditionally, when a linear model is used for the coefficients with respect to the parameters, an analytical solution for the coefficients can be derived directly from the minimization process, often linking them to the PCA eigenvalues and their associated energy content (Jolliffe, 1986). In contrast, the nonlinear nature of the ansatz functions introduced by the input space transformation in our approach necessitates a numerical optimization of $C^{p}_{m,j}$. The input space transformation serves to improve model accuracy by addressing distortions in the parameter space, ensuring a better fit of the surrogate model to the data. Although this approach may alter the conventional interpretation of the coefficients with respect to the principal modes' energy content, it remains a valid and practical method for enhancing surrogate modeling performance. Further methodological details and implications are discussed in our previous work (Fischer and Proppe, 2023).