We gratefully thank both reviewers for their thorough effort in reading the manuscript and providing valuable comments. Point-by-point responses to the reviewers' questions are given below, color-coded for ease of reading. Text in **black** constitutes the original comments by the reviewers, whereas our responses are marked **red,** and suggested changes to the manuscript indicated in **blue**. **When page, section, figure, and line numbers are referenced, we refer to the original, old version of the manuscript.**

In addition to the changes suggested by the reviewers, we noticed we had used a shorter size range than intended when forming the accumulation subrange from the SALSA bins. The upper limit of the SALSA bin 2a3 is 362 nm, not 700 nm, which would instead correspond to bin 2a4 (see Table 1 in Bergman *et al.* (2012)). We extended the range to include bins 2a4 and 2b4, which only had a minor effect on the ECHAM metrics in Figure 3 (in parentheses), as well as the green lines in Figures 2 and 3 (though the difference is too small to notice). This is because only a small fraction of the simulated accumulation subrange particles reside in the 362 – 700 nm range, compared to the 50 – 362 nm range. The main downscaling results were not affected, as the subranges formed from SALSA bins were only used for comparison and not for model training.

Additionally, we have made a few minor changes to the wording of some sentences in the manuscript, as well as some technical corrections to the list of references.

Reply to Anonymous referee #1

This work explores the potential of machine learning techniques in enhancing the accuracy of a global aerosol-climate model's outputs through statistical downscaling. The study focuses on the particle number size distributions from ECHAM-HAMMOZ and data from three European measurement stations were used for downscaling. The results show an improvement in prediction accuracy compared to the original global model outputs. It is a complex and extended study and the results fit within the scope of AMT, being of interest for the international research community. However, I would suggest some aspects to be considered in order to improve the manuscript and/or strengthen its impact before it is published in AMT.

We thank the reviewer for these positive comments regarding our manuscript, and for the suggestions given for improvement.

General – The authors state that the ML methods would certainly improve the spatial accuracy of PNSD derived from models. However, in this study, the ML methods are only applied to specific measurement sites. The models are trained using measurements and global model data, but the horizontal resolution of the climate model is 1.9° × 1.9°, which corresponds to approximately 150 km. So the question is: how would the methods perform at other locations within the same grid cell where no measurement data are available for training? For example, in both Helsinki and Leipzig, there are at least two stations measuring PNSD (urban and traffic). I suggest comparing the model performance at two nearby sites, or at least clarifying the intended utility and applicability of the methods and results presented. From the reviewer's perspective, based on the results shown in this manuscript, these methods do not necessarily improve the spatial accuracy of the model but rather enhance the model's ability to reproduce observations at specific measurement locations.

We are thankful to the reviewer for pointing out to us that the wording of the manuscript could indicate to the reader that our downscaling models increase the resolution of the global-scale model on a more general level; it would be reasonable to assume so, as this is what downscaling methods often aim to do. In our study, corrections to the predicted PNC were indeed only done for each site separately, which we have considered a form of downscaling from the large grid cell area to the point represented by a measurement site. As opposed to a large-scale resolution improvement, the intent of our study is to demonstrate the ML-based statistical downscaling methodology on single stations, which could be extended on in the future by training on more stations and e.g. interpolating between them. This is not to say that the results cannot be valuable without future extension – our methodology can be applied to any site of interest (given sufficient PNSD data) to study the evolution of UFP concentrations in e.g. different climate change scenarios or in the past when observations were not yet available. To clarify the applicability of our results, we have done the following modifications to the manuscript:
L5: changed the sentence to "This study explores the potential of machine learning (ML) techniques in enhancing the accuracy of a global aerosol-climate model's outputs through statistical downscaling to better represent observed data **at specific sites**".
L9: added sentence "A separate ML model was trained for each of the sites and size ranges."
L536: changed the sentence to "This study provides a proof of concept for using ML methods to improve the **site-specific** accuracy of aerosol particle number size distributions derived from global-scale climate models."

Furthermore, the reviewer raised concerns over the generalizability of our ML models to other sites than the ones they have been trained on. We agree it is unlikely that the models would work very well at other sites without further training, as even relatively nearby sites could differ greatly in terms of the features affecting the local air quality (e.g., direction of emission sources, topography of buildings, etc.). However, it was not our intention to apply the trained ML models to a larger area surrounding the stations, but rather apply the correction station-wise, as explained above. Therefore, we deemed further analysis of spatial generalizability not relevant for our intended application.

L130 – The authors defined the nucleation, Aitken, and accumulation mode size ranges using uncommon values (<7.7 nm, 7.7–50 nm, and 50–700 nm). This choice appears to be driven by the bin structure of the SALSA model. If that is the case, I would suggest either avoiding the use of the terms "nucleation," "Aitken," and "accumulation" throughout the manuscript, OR adjusting the size ranges to align with the commonly accepted definitions associated with different aerosol processes (e.g., <25 nm, 25–100 nm, and 100–1000 nm). I would also suggest rephrasing: “These size ranges correspond to the SALSA bins…” by “These size ranges were selected to correspond to the SALSA bins…”

We thank the reviewer for the suggestion; the reviewer is right that the size ranges here differ slightly from the typical size ranges used in the aerosol measurement community. We decided to alter the current naming of the size modes, and in addition, clarify the definition as suggested by the reviewer. We have altered the text starting from line 131 to “These size ranges **were selected to** correspond to the SALSA bins 1a1 for nucleation, 1a2–1a3 for Aitken, and 2a1–2a4 for accumulation"

In terms of the naming scheme, we have now aimed to avoid the word “mode” in order to ensure that our definition does not get as easily confused with the conventional definition of the modes. Instead, we have opted to use the term “subrange” (or, in some cases, other similar words, such as “range”, “sizes”, or “size range”). This change has been applied to all instances of the word “mode” in the manuscript. The only exception is when other studies dealing with modes are referenced. Additionally, when the term is first used on L88, we have added the following clarification:
“We have opted to avoid calling these subranges “modes”, as the subranges do not exactly match the conventional mode definitions due to limitations in the size resolution of the climate model representation.”

Introduction – The first paragraph of the introduction is unclear regarding the distinction between particle number concentrations and mass concentrations. The two terms appear to be used interchangeably or without clear differentiation. I recommend that the authors clarify when they are referring to number concentrations versus mass concentrations and ensure consistent use of these terms throughout the paragraph. It is important to remain that while UFPs mainly control ambient particle concentrations in terms of number, coarser particles control particle concentrations in terms of mass (PM10 and PM2.5).

We agree with the reviewer that discussing both UFPs and PM2.5 (as well as PM10) in the same paragraph can be misleading, as UFPs are much less relevant when considering the mass concentration. Following the reviewer's suggestion, we have added the following disclaimer on L25:

"Different sized particles contribute to different aspects of the ambient particle concentration – UFPs mainly control the concentrations in terms of number, while coarser particles control the concentrations in terms of mass ($PM_{2.5}$)."

Additionally, we have specified in the introduction whenever we are speaking about number concentrations.

L123-129 – Are the PNSDs measured in Germany obtained using a DMPS or a scanning instrument?

We had used an older reference (Hamed *et al.*, 2010) for the Melpitz station, which still listed a DMPS as the measurement instrument. The reviewer is correct in questioning this – a newer publication (Birmili *et al*., 2016), already used as reference for the Leipzig station, also mentions that the Melpitz station utilizes an SMPS instrument. We have therefore modified the manuscript as follows:

L126: changed "DMPS device" to "DMPS/SMPS (Differential/Scanning Mobility Particle Sizer) instruments"

L143: added the sentence "Particle number size distribution is measured with SMPS (TSI) with size range of 5 nm to 800 nm."

What size ranges does each instrument cover? The comparison with the model would be site-dependent if the size distributions differ in their lower and upper diameter limits. Uncertainties of the measurements are not considered?

We have added information concerning the size ranges and names of the instruments used at each site. Melpitz was already mentioned in response to the above comment; for Helsinki and Leipzig, the following changes were made:

L139: added the sentence "In Helsinki, the particle number size distribution is measured with DMPS (TSI), with size range of 3 nm to 1 µm."

L146: modified “At the Leipzig station, the particle size distribution started at 10 nm, …”
to
“The measured particle size range was between 10 nm and 800 nm (using DMPS)”

As the reviewer mentioned, the differences in size distributions might affect the results of the fitting process. This, besides other factors, such as the representativeness of the model grid cell for the measuring station, suggests that the modeling process should be done separately for each site, i.e. the process is site-dependent. In this study, the focus was on verifying how accurate estimates of the number concentrations of different UFP size ranges can be obtained with current data from climate model runs. We didn’t consider the uncertainty related to the measurement data, i.e. the number concentrations reported from the sites were used as “the ground truth” concentrations in training, validation, and test datasets.

Structure – I suggest reconsidering the structure of the sections. For example, the results of the best-performing method are presented in Section 5.1 before the performance of all methods is discussed in Section 5.2. It may be more logical to first present the comparison across all methods, followed by a deeper look at the best-performing one.

Both reviewers suggested this change, and we agree that the revised order is more logical. Accordingly, we have reordered the sections in question.

Additionally, the title of Section 2, "Climate simulation," may not be the most appropriate for the modelling setup. I would suggest something more descriptive, such as "Global model simulations”.
We thank the reviewer for the suggestion. On further thought, we realize that “Climate simulation” may be misunderstood e.g. as a longer simulation than the three years we have used. We have renamed the section “Global aerosol-climate model simulation” to make the title more specific.

Nucleation range differences – In several instances, the authors suggest or conclude that “the nucleation mode proved more challenging to downscale due to high spatial variability and limitations in the underlying large-scale climate model output”. From the reviewer’s perspective, the Aitken mode could also exhibit substantial variability, particularly due to urban emissions. Therefore, a more plausible explanation for the difficulty in downscaling the nucleation mode may lie in the limitations of global models in representing new particle formation (such as the treatment of organics, nitrates, sulfuric acid, or nucleation schemes) rather than primarily in the spatial variability of the sources.

This is an important point. We agree that global climate models do not always represent new particle formation sufficiently. In addition, as Figure 3 indicates, ECHAM-SALSA may potentially have too strong NPF under certain atmospheric conditions, which can lead to overestimation of the number of nucleation subrange particles. We have therefore added to line 380 the following:

"The representation of new particle formation and nucleation–sized particles is, on many occasions, not sufficient in global climate models (Williamson *et al.*, 2019). This can be due to, for instance, errors in estimating nucleation rates. As a study by Laakso *et al*. (2022) shows, ECHAM-SALSA tends to favor partitioning of sulfuric acid to the particle phase due to nucleation over condensation, which may lead to overestimation of nucleation subrange particles. Kokkola *et al.* (2018) compared ECHAM-SALSA number size distributions to observation data, and their results revealed that at some measurement stations, ECHAM-SALSA overestimates the nucleation mode number concentrations. Furthermore, ECHAM-SALSA does not model new particle formation due to nitrates, which may cause differences between modelled and measured nucleation subrange number concentrations. The representation of nucleation-sized aerosols could be enhanced by including a volatility basis set (VBS) scheme (Donahue *et al*., 2011), which can improve the representation of secondary organic aerosols."

Technical corrections

L275 – what means the "-" at the end of the reference?

We removed the dash (which was included in the citation given on the GitHub page) and also added the date of last access.

L130-131 – change "These size ranges correspond to the SALSA bins…" by "These size ranges were selected to correspond to the SALSA bins…"

As mentioned previously, we have now made this change.

L280 - should σ(·) and μ(·) be σ(x) and μ(x)? Actually "x" (eq. 1) is not defined.

We thank the reviewer for noticing this discrepancy. We have removed the placeholder dots and defined the argument "x" of the UCB function on L280:

"…where κ controls how much weight should be given to the posterior's standard deviation **σ(x)** relative to its mean **μ(x) at some point x of the hyperparameter space**"

References:

Bergman, T. ,Kerminen, V.-M., Korhonen, H., Lehtinen, K. J., Makkonen, R., Arola, A., Mielonen, T., Romakkaniemi, S., Kulmala, M., and Kokkola, H.: Evaluation of the sectional aerosol microphysics module SALSA implementation in ECHAM5-HAM aerosol-climate model., Geosci. Model Dev., 5, 845–868, https://doi.org/10.5194/gmd-5-845-2012, 2012.

Birmili, W., Weinhold, K., Rasch, F., Sonntag, A., Sun, J., Merkel, M., Wiedensohler, A., Bastian, S., Schladitz, A., Löschau, G., Cyrys, J., Pitz, M., Gu, J., Kusch, T., Flentje, H., Quass, U., Kaminski, H., Kuhlbusch, T. A. J., Meinhardt, F., Schwerin, A., Bath, O., Ries, L., Gerwig, H., Wirtz, K., and Fiebig, M.: Long-term observations of tropospheric particle number size distributions and equivalent black carbon mass concentrations in the German Ultrafine Aerosol Network (GUAN), Earth System Science Data, 8, 355–382, https://doi.org/10.5194/essd-8-355-2016, 2016.

Donahue, N. M., Epstein, S. A., Pandis, S. N., and Robinson, A. L.: A two-dimensional volatility basis set: 1. organic-aerosol mixing thermodynamics, Atmos. Chem. Phys., 11, 3303–3318, https://doi.org/10.5194/acp-11-3303-2011, 2011.

Hamed, A., Birmili, W., Joutsensaari, J., Mikkonen, S., Asmi, A., Wehner, B., Spindler, G., Jaatinen, A., Wiedensohler, A., Korhonen, H., Lehtinen, K. E. J., and Laaksonen, A.: Changes in the production rate of secondary aerosol particles in Central Europe in view of decreasing SO2 emissions between 1996 and 2006, Atmospheric Chemistry and Physics, 10, 1071–1091, https://doi.org/10.5194/acp-10-1071-2010, 2010.

Kokkola, H., Kühn, T., Laakso, A., Bergman, T., Lehtinen, K. E. J., Mielonen, T., Arola, A., Stadtler, S., Korhonen, H., Ferrachat, S., Lohmann, U., Neubauer, D., Tegen, I., Siegenthaler-Le Drian, C., Schultz, M. G., Bey, I., Stier, P., Daskalakis, N., Heald, C. L., and Romakkaniemi, S.: SALSA2.0: The sectional aerosol module of the aerosol–chemistry–climate model ECHAM6.3.0-HAM2.3-MOZ1.0, Geosci. Model Dev., 11, 3833–3863, https://doi.org/10.5194/gmd-11-3833-2018, 2018.

Laakso, A., Niemeier, U., Visioni, D., Tilmes, S., and Kokkola, H.: Dependency of the impacts of geoengineering on the stratospheric sulfur injection strategy – Part 1: Intercomparison of modal and sectional aerosol modules, Atmos. Chem. Phys., 22, 93–118, https://doi.org/10.5194/acp-22-93-2022, 2022.

Williamson, C.J., Kupc, A., Axisa, D. *et al:* A large source of cloud condensation nuclei from new particle formation in the tropics, *Nature,* **574**, 399–403, https://doi.org/10.1038/s41586-019-1638-9, 2019.