Dear editor and anonymous referees,

Thank you once again for providing us with your insight into our manuscript and suggestions for its possible improvements.

Firstly, we would like to state that as both referees asked about correlations with measurements, we performed a few tests on our data and found a minor error in our code. This meant recalculating our datasets and publishing an updated version to Zenodo along with the corrected python code. Despite minor changes to the validation statistics, the specific qualitative conclusions did not change at all. We apologize for any confusion caused by an error on our side, and we hope to continue the review process with the updated error-free version. Thus, allow us to first briefly address the error in question:

After thoroughly inspecting our code, we realized that the selection of 3-month long moving seasons was performed wrong, i.e., the last day of each such window was mistakenly omitted. The statistical operations on the modeled time series were thus performed on incomplete datasets in both the cross-validation and correction phase of the study, while the final validation was performed using the correctly selected station time series and wrongly selected model data. This did not have any notable effect on the performance of the correction in terms of cross-validation or the spatial distributions of values, but the validation itself (old Fig. 5) was changed more notably. For this reason, the discussion in the manuscript of the old Fig. 5 is changed a little in terms of correlation and NMB, but the other conclusions remained unchanged.

Thank you for your patience regarding errors in the original version of our manuscript. Below, we respond to the referees' remarks to the best of our abilities. Please note that in the light of our recent findings, some of the responses may differ from those previously posted in the interactive discussion. Quotations of the anonymous referees are marked as **R** and our corresponding answer as **A**.

Response to RC1

R: This is a well written paper of good scientific quality and introduces a . The methods are well presented and easy to follow, while the results include all the analysis that would be expected for this type of work. However, some caution should be given in the ability to abstract the results to future climate projections as the characteristics of model quantile biases and their spatial patterns in the present day may be altered under future climate scenarios. This is however a minor outcome of the presented work with the methods used could also be applicable to other present-day uses. I would therefore recommend minor revisions prior to publication.

A: We are pleased to hear that you find our manuscript worthy of publication with minor revisions, and we thank you for your remarks. Below, we address your valuable comments as available in the public discussion.

R: 1) Lines 170-173: The authors appear to be removing very low hourly concentrations from observations: it is not uncommon for ozone levels to be zero or close to zero overnight – these are not unrealistic and therefore should not be being removed purely for quality control reasons. Similarly, there is no removal of high values, leading to an artificial increase in the seasonal mean following the removal of these values. Please can the authors justify the reasons for removing low, but not high, hourly values.

A: We acknowledge the plausibility of measuring very low ozone concentrations at various station sites, especially at night-time hours. Firstly, we would like to clarify that we did not remove low values, but rather the stations measuring them, thus avoiding the artificial increase in the seasonal mean you mention. The goal was to select only stations suitable for our purposes, i.e., stations the character of which would correspond to the central European rural background plausibly captured by the model of 9 km horizontal resolution. Secondly, if a station measures low values too often, it is an indicator of either (i) a close proximity of a source of NOx emissions (NOx-limited regime occurs more naturally in central Europe although we are aware that not necessarily exclusively) or (ii) some form of a local effect impossible to resolve by our simulations (effect of local fine orography, specific fine scale circulation patterns, etc.). In the former case, the station may had been classified as rural background and such a classification did not apply anymore in our studied period, while in the latter case we would be penalizing our simulations in the validation phase for its resolution and inflict the simulations with local errors in the correction phase. Lastly, the choice of stations is based on the choice of Karlický et al. (2024), thus making the presented

manuscript consistent with a previous study. We applied minor changes at the end of section 2.4 to improve clarity of the selection procedure.

R: 2) Line 167 + Lines 173-175: It appears the authors started with 165 stations (line 167), removed some stations (line 174), but then ended with the same number of 165 stations (line 175). Please can these numbers be checked, or clarified as to what process has been undertaken to limit to 165 stations – are all the starting stations still used?

A: We apologize for the misunderstanding. Out of all EEA stations in our domain, 298 stations passed the test of having at least 50% of valid values within the period. Then, 27 stations were removed due to being unsuitable for model validation for aforementioned reasons (thus 271 remained) and the remainder was then iteratively reduced according to the relative position to obtain the final number of 165 stations. This is consistent with Karlický et al. (2024), which considered the exact same choice of stations. We added a better clarification of the station selection into the manuscript from new line 170 onwards.

R: 3) Line 181: The value of α varies monthly. What impact would this have on the results given the choice to look at daily MDA8 values, in particular around the change of months (e.g. 31st compared to 1st of a month), which would presumably impact the predicted values considerably?

A: Generally, in climate model oriented studies, it is considered a standard to perform analyses and bias correction per climatological seasons, i.e., 3 months long periods, thus making our choice of monthly corrections already above the general standard. We are also afraid that 10 years would not be enough for the correction to be performed in higher temporal resolution than 1 month. Discontinuous jumps around the change of months thus present inevitably possible artifacts which are luckily less notable than they would have been if the corrections were made seasonally (as done by, e.g., Rieder et al., 2015). On the other hand, as seen in Fig. 3, the minima of HD for PIQB gauss in particular are not so sharp and so the result is less sensitive to the choice of the specific value of the parameter. We added a more thorough justification (new lines 186—197) for our choice of parameters.

R: 4) Figure 9,10: It may be helpful to guide the reader by distinguishing between exceedances of observations and of the simulation contours in different colours (e.g. use a different outline colour for observations).

A: Such a change is certainly possible. We changed the color scheme of the figures, including a different outline color for observations. We also double-checked the updated color scheme for color-blindness tests.

R: 5) Lines 433-434: How confident are the authors that the systematic errors of the models in the present-day climate would still exist under future climate projections, or will model quantile bias characteristics and spatial variabilities change over time?

A: In our opinion, this question concerns the limitations of climate modeling in general. In other studies, this is typically an unspoken implicit assumption, which we explicitly stated in the introduction (line 37). We are very confident that bias correction can reliably mitigate present-day systematic errors originating from model resolution (e.g., uncertainties regarding emission inputs, boundary conditions, etc.) and errors of this type are bound to occur in corresponding projections as well, regardless of the definition of an error (e.g., quantile bias, mean bias, etc.). However, as also stated in the introduction (line 41; Liu et al., 2022), the bias patterns may differ depending on the chemical processes taking place. To our knowledge, a bias correction strategy involving the individual model processes has not yet been introduced, therefore, one must rely on corrections of the type which follows some definition of an error, which were compared by, e.g., Staehle et al. (2024). For completeness, we may state that the reliability of the correction should be high for similar climate conditions (e.g., near future projections) and reliable only in terms of resolution-induced deficiencies and otherwise potentially unreliable for projections in distant future (a trait shared with any other concurrent bias correction strategy due to potentially different bias patterns). We modified the phrasing of the last sentence in the conclusion of our manuscript.

Additionally, we thank you for notifying us on several technical errors:

R: Line 156: “spatial extend depicted” should be "spatial extent depicted"

A: Corrected.

R: Line 216: "Tab. 1.” should be Table 1. (See https://publications.copernicus.org/for_authors/manuscript_preparation.html#figurestable

A: Corrected on former line 216 (new line 230) as well as on former line 277 (new line 294).

R: Line 331: “nort-west” should be "north-west"

A: Corrected.

R: Figures 9, 10: These figures should be checked for colour-blindness as red and green are commonly confused.

A: We changed the color scheme of these figures.

References:

Karlický, J., Rieder, H. E., Huszár, P., Peiker, J., and Sukhodolov, T.: A cautious note advocating the use of ensembles of models and driving data in modeling of regional ozone burdens, Air Quality, Atmosphere & Health, 17, 1415–1424, https://doi.org/10.1007/s11869-024-01516-3, 2024.

Liu, Z., Doherty, R. M., Wild, O., O'Connor, F. M., and Turnock, S. T.: Correcting ozone biases in a global chemistry–climate model: implications for future ozone, Atmospheric Chemistry and Physics, 22, 12 543–12 557, https://doi.org/10.5194/acp-22-12543-2022, 2022.

Rieder, H. E., Fiore, A. M., Horowitz, L. W., and Naik, V.: Projecting policy-relevant metrics for high summertime ozone pollution events over the eastern United States due to climate and emission changes during the 21st century, Journal of Geophysical Research: Atmospheres, 120, 784–800, https://doi.org/https://doi.org/10.1002/2014JD022303, 2015.

Staehle, C., Rieder, H. E., Fiore, A. M., and Schnell, J. L.: Technical note: An assessment of the performance of statistical bias correction techniques for global chemistry–climate model surface ozone fields, Atmospheric Chemistry and Physics, 24, 5953–5969, https://doi.org/10.5194/acp-24-5953-2024, 2024.

Response to RC2

R: The manuscript introduces the Parametric Interpolation of Quantile Biases (PIQB) as a novel strategy for ozone bias correction in high-resolution simulations. While the methodology is promising, several areas regarding temporal continuity, statistical reliability for policy application, and the depth of spatial analysis require further clarification and improvement.

A: Thank you for your suggestions for the improvement of our manuscript. Below, we address your comments and the way we hopefully enhanced the corresponding parts of our manuscript.

R: 1) The authors optimize the interpolation parameter on a discrete monthly basis. While this captures seasonal shifts, it risks introducing artificial "jumps" or discontinuities at the boundaries of each month. Since the study already utilizes a 3-month "moving season" for cross-validation, it is unclear why a similar sliding window was not applied to the evolution of to ensure temporal smoothness.

A: In general climatology, it is considered a standard practice to evaluate the model performance, conduct bias correction and postprocessing, etc., on a discrete seasonal basis. We opted for a monthly basis to increase the temporal resolution of our corrections precisely to capture shifts also at the boundaries of the seasons. The 3-month moving season was introduced to prevent overfitting, since performing purely monthly bias correction on 10-year long simulations may introduce statistical errors. Regarding temporal smoothness, we find the annual cycles of the optimal parameters to be already smooth enough for climatological purposes and the residual errors in the current Fig. 5 show a clear annual cycle as well. We added a more in-depth justification for the 3-month moving window into section 2.5 in the manuscript on new lines 186—197.

R: 2) Figure 5 shows a significant decrease in the Pearson correlation coefficient after correction. The authors attribute this to the "shuffling" or permutation of data when using neighboring months for calibration. For policymakers, the timing of peak ozone events is as critical as the absolute magnitude. If the model loses its ability to capture when pollution events occur, the reliability of the correction for real-world health alerts is compromised. The authors should discuss the implications of this "temporal de-correlation" in more detail.

A: As stated at the beginning of this letter, the spatio-temporal correlations were found to be increased after the correction, which we attribute mostly to the better resolved spatial

distribution of the simulated MDA8 and we apologize for our previous error. On the other hand, as we stated in the interactive discussion, our model experiments were designed in such a way that the meteorological and chemical boundary conditions were not guaranteed to share the same weather conditions, meaning that the temporal correlation is not a reliable metric and we only wish to address the plausible spans of values before and after correction, thus making other metrics much more relevant. However, the spatial component of correlation may help to explain the differences between the strategies. For these reasons, we included a brief explanation of the correlation improvements to section 3.3.1 (new lines 306—316) and a further emphasis on the limited importance of the temporal component of the correlations (new lines 240—242).

R: 3) The text frequently mentions that traditional methods (like Adjoint PDFs) introduce statistical artifacts. However, the results section lacks a direct, high-contrast visual comparison that explicitly highlights these artifacts versus the PIQB results. Adding zoomed-in panels for complex terrain (e.g., the Alps) would better substantiate the claim that PIQB avoids these pitfalls.

A: We believe that zoomed-in panels would not provide the readers with any new information. Instead, we suggest displaying fields of quantile biases as predicted by each strategy for the quantiles of 5, 50 and 95 (i.e., subtracting the corrected quantiles from the original quantiles). This way, it can be shown that, e.g., Adjoint PDFs lower the already underestimated central European MDA8 while overshooting the already overestimated median MDA8 in the Po Valley. We included such figures for both the WRF-Chem and the CAMx simulations in the summer period in the appendix, and we added references to these figures in the discussion in sections 3.3.2 and 3.3.3.

R: 4) The core of the PIQB strategy relies on a hybrid formulation (Eq. 4–7). For readers with a non-mathematical background, a flowchart or conceptual diagram illustrating how "model support" and "station quantile biases" are fused would greatly enhance the paper’s accessibility.

A: We understand the issue of accessibility, and we included such a figure in the manuscript. The chart displays 3 steps, the left one shows the definition of quantile bias, the middle one shows the difference between interpolating station data and station quantile biases, and lastly the right figure demonstrates the optimization procedure. We believe that readers of various backgrounds should now be able to picture the process.

R: 5) In Section 3.3, the discussion of spatial gradients (zonal/meridional) in Figures 7 and 8 is largely qualitative. While the authors note that Obs. IDW "completely smooths out" variability, they should provide quantitative metrics—such as spatial correlation coefficients or spatial RMSE—to rigorously compare the "Spatial Integrity" of PIQB against the other strategies.

A: We provided a partly quantitative comparison in section 3.2 which compares the ability of each strategy to interpolate the station information into the regular grid, and additionally few remarks regarding spatial correlation into section 3.3.1. Regarding comparing individual strategies, we are afraid that comparing them to PIQB (with whichever interpolator) would only confuse the reader, since we do not know what the “ground truth” mean MDA8 field looks like and doing so could result in a misleading conclusion that we claim PIQB to provide the actual MDA8 fields. This is the main reason for regressing to qualitative discussions as there is no direct quantity to be compared. On the other hand, as a response to your remark 3), we added figures with quantile biases into the manuscript’s appendix, which should demonstrate the characteristics of the specific strategies mentioned in the discussion of section 3.3.2.

R: 6) The current Section 3.3 is quite dense and covers multiple validation dimensions simultaneously. To improve readability and logical flow, I recommend subdividing this section into the following thematic subsections: 1） Statistical Fidelity: Focus on , NMB, and PDF matching at station sites ；2） Temporal Dynamics: Analyze the annual cycle and the impact of correction on temporal correlation ；3） Spatial Integrity: Quantitatively evaluate the preservation of spatial gradients and model-resolved features. 4） Policy-Relevant Metrics: Focus on the exceedance days (MDA8) and the success rates of the confusion matrix.

A: We agree to subdivide the current section 3.3 into subsections, although into 3 instead of 4. As explained above, we do not find the discussion of “temporal dynamics” in regard to correlation to be fruitful, and so we consequently moved the former Fig. 6 to the supplement, as it is unnecessary to have a subsection discussing a single figure. Furthermore, as also stated above, fully quantitative discussions are not possible for certain aspects of our work, or at least not in the fashion you suggest in one of your previous remarks - we consider it important to show comparisons which do not apriori highlight any of the presented strategies to retain objectivity. Other than that, we agree with the suggested overall layout,

and we divided the section into 3.3.1 Statistical validation on station data, 3.3.2 Spatial distributions of MDA8 and 3.3.3 Immission limit exceedances.

We hope to have replied in a satisfactory enough manner to both anonymous referees, and we look forward to continuing in the review process.

Kind regards,

Jan Peiker & the co-authors.