https://doi.org/10.5194/egusphere-2023-2743
Preprint. Discussion started: 23 November 2023

Technical note: An assessment of the performance of statistical bias correction techniques for global chemistry-climate model surface ozone fields

Christoph Staehle[1], Harald E. Rieder[1], Arlene M. Fiore[2]

[1]Institute of Meteorology and Climatology, University of Natural Resources and Life Sciences, Vienna, Austria

[2]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

Correspondence to: christoph.staehle@boku.ac.at

Abstract. State of the art chemistry-climate models (CCMs) still show biases compared to ground level ozone observations, illustrating remaining difficulties and challenges in the simulation of atmospheric processes governing ozone production and loss. Therefore, CCM output is frequently bias-corrected in studies seeking to explore changing air quality burdens and associated impacts. Here we assess four statistical bias correction techniques of varying complexity, and their application to surface ozone fields of four CCMs, and evaluate their performance against gridded observations in the EU and US. For the evaluation of the raw CCM outputs and the performance of the individual adjustment techniques we focus on two time periods (2005-2009 & 2010-2014), where the first period is used for development and training and the second to evaluate the performance of techniques when applied to model projections. Our results show, that while all methods applied are capable of significantly reducing the model bias, better results are obtained for more complex approaches such as quantile-mapping and delta-functions. We also highlight the sensitivity of the correction techniques to individual CCM skill at reproducing the observed distributional change in surface ozone. Ensemble simulations available for one CCM indicate the ozone bias arises from sensitivities in chemical mechanisms or emissions rather than driving meteorology.

1. Introduction

Surface ozone (O_3) is both an air pollutant and greenhouse gas, formed in photochemical reactions involving precursor substances such as nitrogen oxides (NO_x) and volatile organic compounds (VOCs) of both anthropogenic and non-anthropogenic origin [e.g. Checa-Garcia et al., 2018; Lelieveld & Dentener, 2000; Monks et al., 2015]. In addition to the availability of precursor gasses, the NO_x to VOC ratio as well as solar radiation and ambient air temperature, controlling emissions of biogenic VOCs (BVOCs) and chemical reaction rates, play a crucial role for O_3 formation [Chameides et al., 1988; Sillman, 1999; Sillman et al., 1990]. O_3 is associated with a variety of detrimental human health effects, especially in the context of the respiratory and cardiovascular system, resulting in about 5-20 % of premature deaths attributable to ambient air pollution [Monks et al., 2015]. In addition to its negative health effects O_3 is considered to compromise the metabolism of plants through stomatal uptake and to cause damage to leaf surfaces, thereby affecting biomass and crop production [Da et al., 2022; EEA, 2020; Fleming et al., 2018; Mills et al., 2018; Monks et al., 2015]. Consequently, a large body of studies examine past, present and future development of surface O_3 burdens as well as resulting health and ecological impacts on both regional and global scale [e.g. Da et al., 2022; Meehl et al., 2018; Westervelt et al., 2019].

Studies seeking to explore future changes in surface O_3 burdens and their implications for human health and the biosphere rely on outputs of chemistry-climate models (CCMs) and chemistry-transport models (CTMs). However, despite ongoing development, these models show deficiencies in the adequate representation of tropospheric O_3 on regional to local scale and changes therein. Especially global scale models are known to be biased compared to observations [e.g. Griffiths et al., 2021; Turnock et al., 2020; Young et al., 2018], raising the question regarding the reliability of the simulated surface ozone response to changes in precursors and ambient climate. The number of possible reasons for the deviation of model output and observations is basically increasing with the complexity of the models. However, literature commonly suggests issues with emissions fed into the models, the applied chemical mechanism, meteorology or deposition in addition to uncertainties associated with the spatial resolution [e.g. Archibald et al., 2020; Liu et al., 2022]. To overcome this issue, similar to climate studies, statistical bias correction techniques of different complexity are frequently applied to correct global model fields. Such corrections allow the diagnosis of changes in ambient meteorological conditions and ozone in isolation or combination and to investigate related impacts on human health. Machine learning approaches are increasingly being used for correction purposes [e.g. Liu et al., 2022]. These methods, however, usually have the disadvantage of behaving like a "black box", i.e. algorithms lack traceability and thus physical insights as to the root cause of biases. To date no detailed comparison of different statistical bias correction techniques for surface ozone burdens has been performed and the present study aims to close this gap.

Here we analyze historical simulations from 3 different global CCMs contributing to the Coupled Model Intercomparison Project phase 6 (CMIP6), as well as a 13-member ensemble of the CESM2-WACCM6 model for the European (EU) and contiguous United States of America (US) domain. For an assessment of model performance, we compare model outputs with gridded observational datasets available for both domains. First, we evaluate the ozone fields of the individual CCMs with observations and contrast the magnitude, sign and seasonality of the bias among CCMs. Thereafter, we apply a set of statistical bias correction techniques aiming for a reduction of the initial bias, independent of its origin, and evaluate the performance of these methods to identify if a particular correction technique is preferable across models.

Since the model simulations are "free-running", and thus create their own meteorology internally, a direct day-to-day comparison with the observations is not meaningful. Hence, our analysis primarily aims to evaluate the distribution of the O_3 fields in a statistical sense. Given the importance of ozone for human health we focus on the upper tail of the maximum daily 8-hour average (MDA8) O_3 distribution, and the frequency of occurrence of exceedance of health-related target values for Europe and the US.

2. Data & Methods

2.1. Model data

The O_3 data sets explored in our analysis are hourly surface O_3 outputs from three CCMs (GFDL-ESM4, UKESM1-0-LL and EC-Earth3) contributing to CMIP6, and a 13-member ensemble simulation created with CESM2-WACCM6 (note only the first member is used for the main part of this study), from which we calculate maximum daily eight-hour average (MDA8) O_3 values. We also use observational MDA8 O_3 data derived from gridded hourly datasets, with a spatial resolution of 1° x 1° per grid cell and a temporal extent of 1993 to 2014 for both the European and the US domain [Schnell et al., 2014].

To allow for an optimal comparison, the model data is regridded using an ordinary inverse distance weighting algorithm to match the spatial extent of the observations. In addition, all data sets are harmonized regarding their temporal resolution by removal of days not included in any of the other datasets, resulting in a 358-day calendar (30 days per month except for February). MDA8 O$_3$ is derived for each dataset and time step. For the historical analysis we use 2005 to 2009 to evaluate the baseline bias of the individual CCMs and establish the performance of individual bias correction techniques. The time slice 2010 to 2014 is used subsequently, to evaluate the performance of our methods for model projections.

2.2. Bias correction methods

For statistical bias correction we apply four different techniques, which are detailed below. Here M_q and O_q denote quantiles ($q \in 1, \ldots, N \mid 1 = min, N = max$) of the model and observational distributions, respectively. The running index j marks individual MDA8 O$_3$ model values. Additionally, we use the indices $hist$ and $proj$ to differentiate between historical and projected data. Primed terms indicate the bias corrected model outputs.

2.2.1. Mean bias correction (MB)

$$\overline{\Delta M} = \frac{1}{N} \sum \left(M_q^{hist} - O_q^{hist} \right) \qquad (1)$$

$$= \overline{M_q^{hist}} - \overline{O_q^{hist}}$$

$$M_q^{\prime proj} = M_q^{proj} - \overline{\Delta M} \; ; \; M_q^{\prime hist} = M_q^{hist} - \overline{\Delta M} \qquad (2)$$

The MB is a commonly used approach assuming a constant offset between model and observations. As an initial step we derive the average difference of the historic model and observational percentiles (alternatively the difference between the mean values of both empirical cumulative distribution functions (ECDFs) can be computed). Subsequently we subtract the result of Eq. (1) from each quantile of the projected model distribution to retrieve a bias corrected model ECDF (Eq. (2)).

2.2.2. Relative bias correction (RB)

$$\bar{c} = \frac{1}{N} \sum \left(\frac{M_q^{hist} - O_q^{hist}}{O_q^{hist}} \right) \qquad (3)$$

$$M_q^{\prime proj} = M_q^{proj} - \bar{c} * O_q^{hist} \qquad (4)$$

Here, similar to the MB method, we assume that model and observations differ by a constant factor. In contrast to the MB correction, however, we derive the average of the relative deviation of the historic model and observational percentiles (Eq. (3)). The bias corrected model projection (Eq. (4)) is then calculated as the difference between the raw model and the observed quantiles times the correction term established in Eq. (3).

2.2.3. Delta correction (DC)

$$\Delta M_q = M_q^{proj} - M_q^{hist} \qquad (5)$$

$$M_q^{\prime proj} = O_q^{hist} + \Delta M_q \qquad (6)$$

The DC approach follows the methodology detailed in Rieder et al. [2018]. In contrast to the MB and DC methods it is assumed that, while the individual model values may be biased, the system response (i.e., change between two time periods) is represented adequately by the model. Therefore the deviation between future and base period model data is calculated for all quantiles individually (Eq. (5)). Finally the corrected model projection is derived as the observed distribution plus the initially computed model change (Eq. (6)).

2.2.4. Quantile mapping (QM)

$$R_q^{hist} = \frac{O_{q+1}^{hist} - O_q^{hist}}{M_{q+1}^{hist} - M_q^{hist}} \tag{7}$$

$$c_j^{hist} = R_q^{hist} * \left(M_j - M_q\right);\ M_q \leq M_j < M_{q+1} \tag{8}$$

$$M_j'^{hist} = O_q^{hist} + c_j^{hist} \tag{9}$$

$$\Delta M_q = M_q^{proj} - M_q^{hist} \tag{10}$$

$$m_q'^{proj} = M_q'^{hist} + \Delta M_q \tag{11}$$

$$R_q^{proj} = \frac{m_{q+1}'^{proj} - m_q'^{proj}}{M_{q+1}^{proj} - M_q^{proj}} \tag{12}$$

$$c_j^{fut} = R_q^{proj} * \left(M_j^{proj} - M_q^{proj}\right);\ M_q^{proj} \leq M_j^{proj} < M_{q+1}^{proj} \tag{13}$$

$$M_j'^{proj} = m_q'^{proj} + c_j^{proj} \tag{14}$$

The term "quantile mapping" summarizes a variety of similar bias correction approaches used within the climate research community [e.g. Lehner et al., 2023]. Here, however, we follow the method described for CCM outputs in Rieder et al. [2015]. In contrast to the other methods used in this study, the QM is a multi-step approach. The first steps, illustrated in Eq. (7) to (9), consist of the computation of a bias corrected historic model distribution. Next, the result is used to create a distribution of a bias corrected future ECDF, similar to the DC method (Eq. (10) and (11)), which is then employed to derive the bias corrected future model data (Eq. (12) to (14)). In contrast to Rieder et al. [2015] however, who suggested a fixed apportionment for the quantiles used to avoid non meaningful results by executing not defined operations, especially in Eq. (7) and (12), (i.e. denominator equals zero or both denominator and numerator equal zero), we employ here a variable algorithm selecting the optimal amount of percentiles for each realization of the QM method individually. This is achieved by fixing the minimum and maximum values of the model ECDF and allowing for all quantiles with unique values within this range, i.e. if several quantiles share the same value, which might be the case, especially for narrow distributions, only the first quantile is used.

All four methods are applied to the ECDFs of the individual CCM datasets (1) on a monthly basis within the base time interval, (2) for each grid cell individually (in contrast to Rieder et al. [2015], who used a regional approach), (3) for both the EU and US domains. While it is implied that the model data differs from the observations by a constant factor for the MB and RB methods, the DC and QM techniques assume that the difference between the future and reference period is represented adequately in the individual models, independent of the prevailing model bias. In contrast to the QM method, which provides the opportunity to directly correct individual daily MDA8 O_3 values, the application of the MB, RB (according to the methodology detailed above) and DC techniques solely results in new model ECDFs. The mapping algorithm, detailed in Eq. (7) to (9) is therefore applied further to the outputs of these three correction methods. Thereby the model data is mapped onto the bias corrected ECDFs, allowing for an optimal comparison of original- and bias corrected model data with the observations and the results from the other correction techniques under investigation here.

To quantify the initial biases as well as the remaining bias after application of the individual correction techniques we derive the number of days above the target value for the protection of human health (120 µg/m^3 in the EU (approximately 60 ppb) and 70 ppb in the US) and the residual bias of the ECDFs on seasonal and annual time scales [EPA, 2015; EUR-LEX, 2011].

3. Results

3.1 Model evaluation

We start by evaluating the performance of the global models in representing the MDA8 O_3 burden for the historical time period (2005-2009). Fig. 1a,b shows the MDA8 O_3 probability density function for models and gridded observations for the EU and US domains. Pronounced differences emerge between the individual models and observations for both domains. Generally, the models are biased high compared to observations, and the amplitude of the bias varies substantially among models. One exception in this regard is the EC-Earth3 model, which shows a high bias compared to observations across the majority of the MDA8 O_3 distribution but in contrast to other models a low bias at the upper tail.
We further investigate the magnitude of the model biases in Fig. 1c,d by contrasting the annual average number (and seasonal partitioning) of days above the target value to protect human health, defined as 60 and 70 ppb for the EU and US domains, respectively. For the observations we find a domain average number of exceedance days of the target values of 8 (five for summer and 3 for spring) and 3 (2 for summer and 1 for spring) days for the EU and US domains in 2005-2009. While the models agree with observations regarding a preferred occurrence of non-attainment days in summer season, all models but EC-Earth3 substantially overestimate the occurrence frequency of exceedance days. The domain average bias in non-attainment days for the EU ranges between 5 days in EC-Earth3 and 113 days in UKESM1-0-LL. In contrast, values for the US vary between 2 to 79 days. Overall our findings indicate a slightly better agreement of CCMs regarding the policy relevant metrics in the US than EU, which has to be taken with caution given also the regional difference in the MDA8 O_3 target value. Assuming the same target threshold as for Europe, we find that the number of exceedance days ranges between 20 and 174. Table 1 provides a summary of the occurrence frequency of MDA8 O_3 extremes for models and observations on an annual and seasonal basis (note, fall and winter are grouped together (FW) due to the small number of exeedance days derived for these seasons).

https://doi.org/10.5194/egusphere-2023-2743
Preprint. Discussion started: 23 November 2023

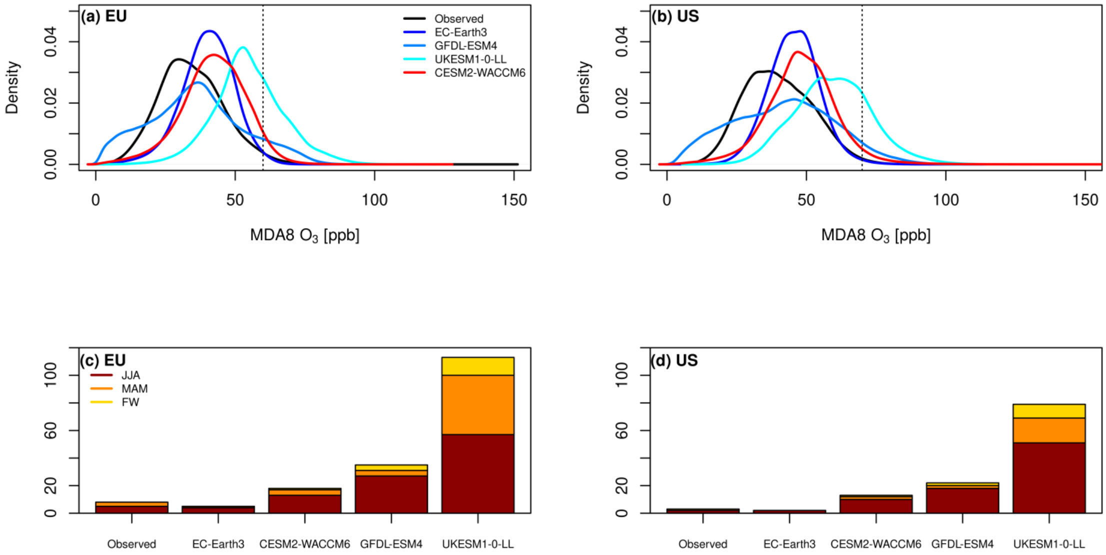
2005-2009					
	Obs	**EC-Earth3**	**CESM2-WACCM6**	**GFDL-ESM4**	**UKESM1-0-LL**
JJA	5 (2)	4 (2; *12*)	13 (10; *29*)	27 (18; *42*)	57 (51; *78*)
MAM	3 (1)	1 (0; *4*)	4 (2; *17*)	4 (2; *10*)	43 (18; *63*)
FW	0 (0)	0 (0; *4*)	1 (1; *5*)	4 (2; *8*)	13 (10; *33*)
annual	8 (3)	5 (2; *20*)	18 (13; *51*)	35 (22; *60*)	113 (79; *174*)
2010-2014					
JJA	3 (1)	1 (1; *6*)	9 (7; *24*)	29 (17; *40*)	54 (37; *72*)
MAM	1 (0)	0 (0; *2*)	3 (1; *15*)	2 (1; *6*)	39 (10; *58*)
FW	0 (0)	0 (0; *2*)	1 (1; *6*)	4 (1; *8*)	14 (6; *27*)
annual	4 (1)	1 (1; *10*)	13 (9; *45*)	35 (19; *54*)	107 (53; *157*)

Table 1: Average number of exeedance days per grid cell derived from observations and individual raw model data for the EU and US (given in parenthesis) for summer (JJA), spring (MAM), fall and winter (FW) and annual. Note numbers in italics in the parentheses are derived applying the EU threshold for the US.

Next, we turn to model biases in the spatial domain. Figure 2 shows the difference in the average number of days above the target value for individual models to observations (note, grey shaded areas indicate a marginal difference of up to ± two days). The spatial distribution of differences confirms the biases detailed above, showing regionally varying but distinct biases of the models examined. Of the models examined, the EC-Earth3 model performs best in both domains with a domain average bias of +7 (EU) and +3 days (US), respectively. While pronounced differences in the magnitude of the bias between individual models occur, the spatial patterns in bias occurrence are quite similar. In particular a north to south gradient emerges in the European domain with significantly higher model biases in the Mediterranean region and small to negligible biases in Scandinavia and the UK. For the US we find across models less pronounced biases in the Midwest, while substantial biases emerge in the North- and Southeast t and Southwest.

To investigate the consistency of the spatial bias in models compared to observations we expand the analysis to the 2010-2014 time period (Fig. S1 & S2). Although slight variations are found for individual seasons, overall the result for this time period resembles those obtained for 2005-2009 in both the US and EU domain (see Fig. S1). This result provides further confidence in the robustness of our assessment of general model biases in the MDA8 O_3 distribution and the modelled frequency of non-attainment days.

https://doi.org/10.5194/egusphere-2023-2743
Preprint. Discussion started: 23 November 2023



190 **Fig. 1:** Probability density function (PDF) of observed (black) and modeled (coloured) MDA8 O_3 during 2005 to 2009 in the EU (a) and US (b) domain. Average number of days above the MDA8 O_3 target value per grid cell for summer (JJA – red), spring (MAM – orange) as well as fall and winter months (FW – yellow) during 2005 to 2009 in the EU (c) and US (d) domains. In panels (a) and (b) dashed vertical lines indicate the target value for the protection of human health. The annual average number of exceedance days in (c) to (d) is given by the sum of the individual segments, i.e. the total height of the
195 bars.

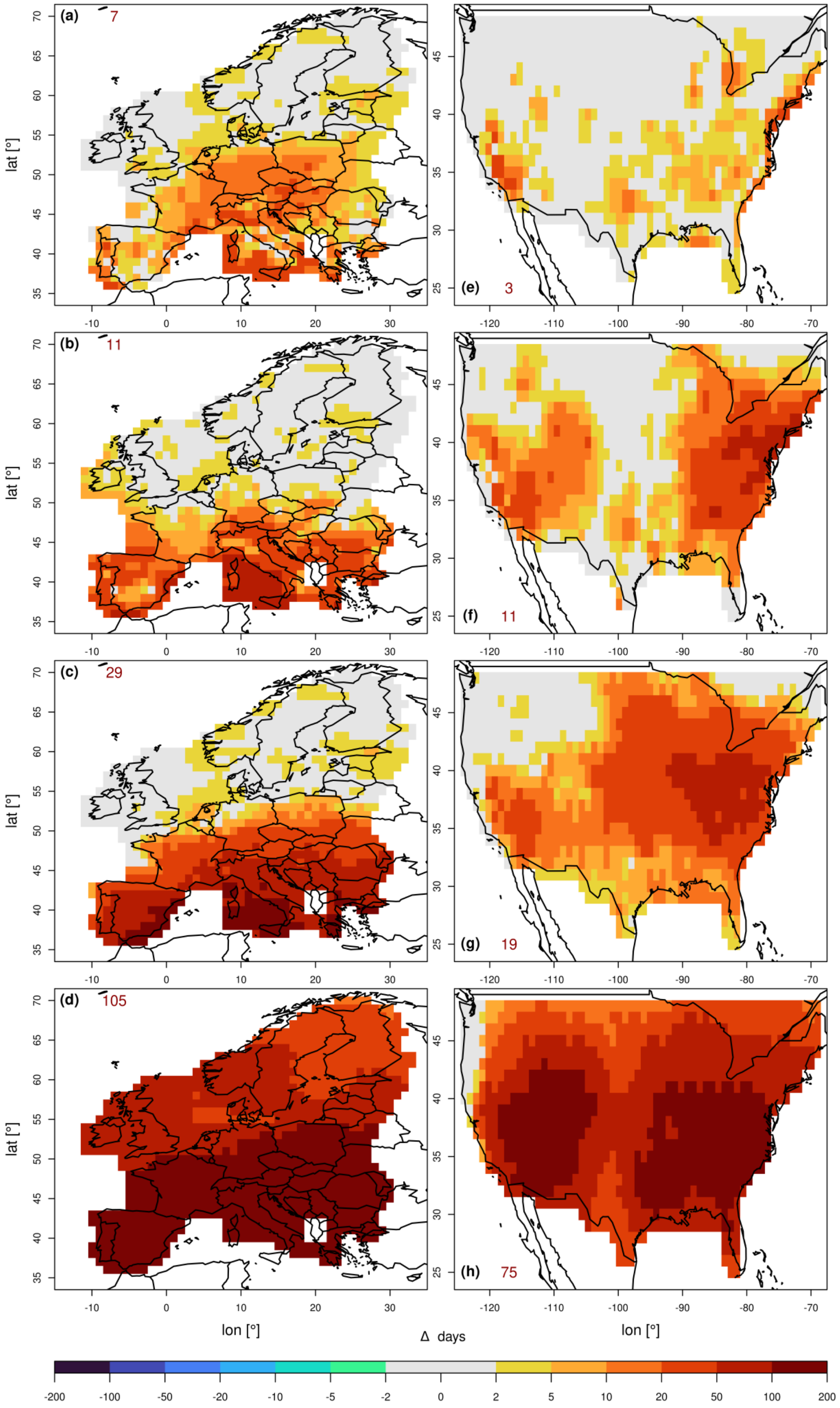

(a) 7
(b) 11
(c) 29
(d) 105
(e) 3
(f) 11
(g) 19
(h) 75
lat [°]
70
65
60
55
50
45
40
35
45
40
35
30
25
-10
0
10
20
30
-120
-110
-100
-90
-80
-70
lon [°]
Δ days
-200
-100
-50
-20
-10
-5
-2
0
2
5
10
20
50
100
200

Fig. 2: Difference in the average number of days above the MDA8 O_3 target value in CCM simulations (EC-Earth3 a & e, CESM2-WACCM6 b & f, GFDL-ESM4 c & g, UKESM1-0-LL d & h) compared to gridded observations for the European (left) and US domain (right). All panels for differences during 2005-2009. Red numbers in the upper/lower left corner indicate the grid cell average anomaly. Grey shading indicates differences within ± 2 exceedance days.

3.2. Bias correction for the base period 2005-2009

Having illustrated the model biases for the past, we turn next to bias correction. To this end we apply the individual bias correction methods to model outputs for 2005-2009 and evaluate their performance for the MDA8O_3 distribution and the number of non-attainment days. The DC method represents an exception in this case as applying this method, by definition, would yield a “perfect” agreement with the observational ECDF. Accordingly, any potential deviations from observed ECDF would be a mere result of uncertainties associated with implementation, in particular the mapping algorithm and rounding, and thereby do not represent the performance of the DC method in context of the base dataset. The performance of the DC method will be however, assessed, along with the other methods when applied to the evaluation period 2010-2014.

Figure 3 shows the distribution of the grid-cell level bias in the number of exeedance days for the European (a – d) and US (e – h) domains. All methods applied reduce the bias substantially. The MB and RB methods yield similar results. Both methods tend to overcorrect the bias, yielding residual biases for individual grid-cells varying between -22 and +8 days (EU) and -10 to +6 days (US), with MB performing slightly better. In contrast, the QM method yields an almost perfect agreement (comparable to the DC method as detailed above) with observations. Residual biases are between -2 and +1 days for Europe and 0 days for the US.

Spatial distributions of the anomaly on exeedance days, are illustrated in Fig. S3 and S4. We find that the application of a particular method yields similar spatial patterns of improvement independent of the model to which it is applied and of the initial model bias. For the MB and RB approaches, the spatial gradient in the bias identified in the raw models remains for the EU domain, although with reversed sign for the majority of applications, i.e. stronger overcorrection in the Central Europe and the Mediterranean than in the northern parts of EU domain. For the US the MB and RB methods perform better compared to Europe. This finding, however, is attributable to the higher target threshold rather than the actual performance of these methods as shown in section 3.1. The QM method corrected outputs agree best with the observations in both domains.

We examine the PDFs of the bias corrected model data for conformity with the observations (see Fig. S5). While all correction methods lower the bias across the whole distribution, the MB and RB approaches, still deviate from the observations. In contrast, the distribution of the QM-corrected data is almost perfectly aligned with the observational PDF, independent of the model and domain. In summary, our evaluation for the baseline period indicates a clear preference for the QM (or the DC) method.

https://doi.org/10.5194/egusphere-2023-2743
Preprint. Discussion started: 23 November 2023

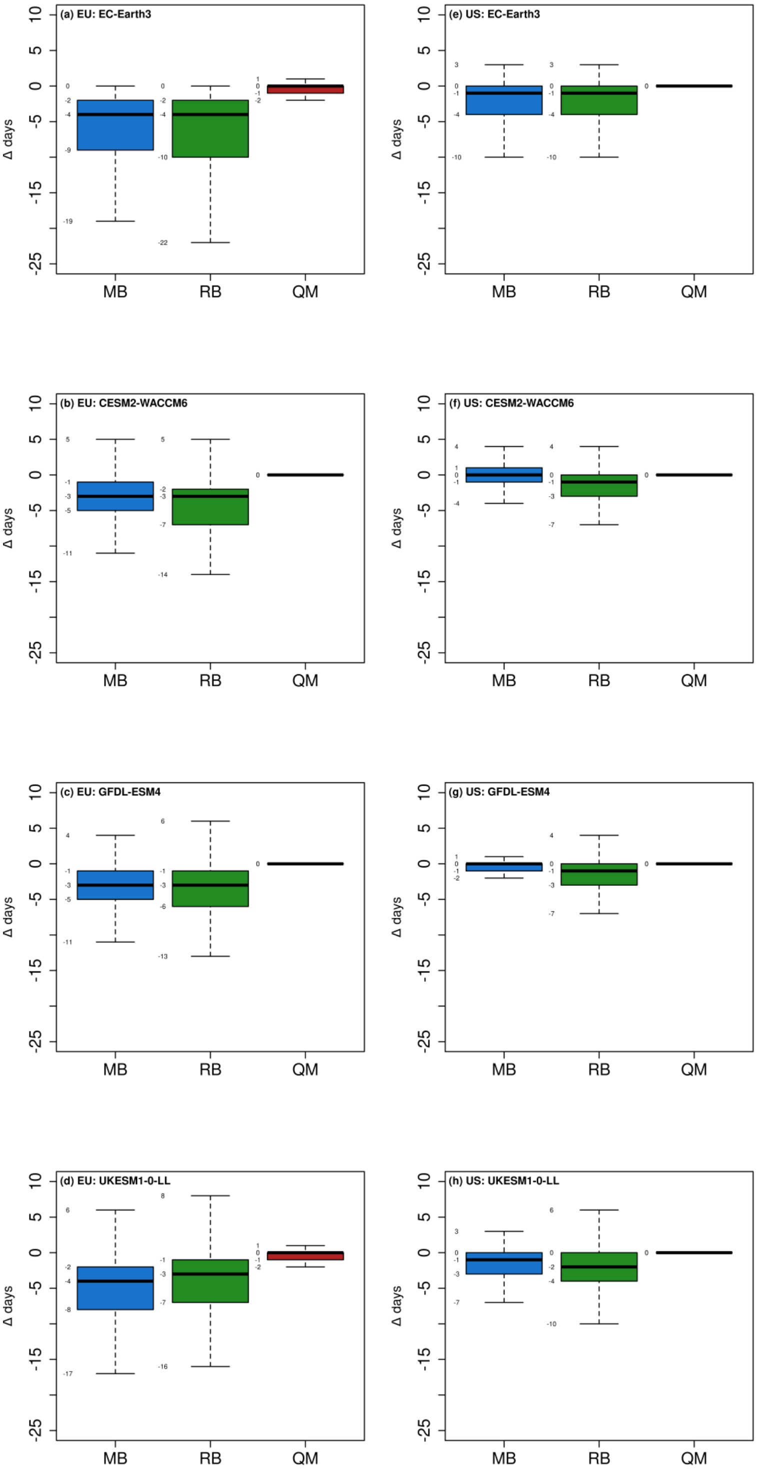


Fig. 3: Boxplots of the average residual bias in exceedance days pooled across grid cells in the individual CCMs in 2005-2009: (a) EC-Earth3, (b) CESM2-WACCM6, (c) GFDL-ESM4 and (d) UKESM1-0-LL models in the EU domain, (e),(f),(g),(h) as (a),(b),(c),(d) but for the US domain. Blue, Green and red colour indicate the MB, RB and QM correction methods respectively.

3.3. Bias correction performance in the evaluation period 2010-2014

Next, we turn the focus to the results obtained with individual bias correction techniques during the evaluation time period (2010-2014). We apply the adjustment methods to the MDA8 O_3 outputs of the individual models, but treat the data as independent realizations, in order to assess the methods performance for their applicability to future projections (see section 2).

Figure 4 shows the distribution of the residual grid-cell level mean bias compared to observations for the number of exceedance days of the target value. Here we find a larger residual bias, ranging between -17 and +11 exceedance days in the European domain (Fig. 4a-d) than in the US (Fig. 4e-h) where the bias after correction varies at grid level between -5 to +5 days. Furthermore, contrasting the performance of the individual bias correction techniques yields a curious result, as we no longer identify an individual correction technique as optimal across models and spatial domains.

We further explore the spatial distribution of the residual bias. Compared to the base period the MB and RB (see Fig. S6-7 a-d & e-h) corrected models show an improved agreement compared to observations. While the spatial patterns of bias distributions are similar to the 2005-2009 period (except for the GFDL-ESM4 model) an improvement compared to the base period is found for the northern and eastern European countries as well as the south-east US. The residual bias worsens in the central EU and the Mediterranean as well as the south-west US when applied to the GFDL-ESM4 model. For the DC and QM approaches (see Fig. S6-7 i-l & m-p) on the other hand, we find a significantly increased residual bias (of both positive and negative sign), independent of model and domain.

Although all methods applied are still capable of significantly reducing the bias, these results, in contrast to those for the base period, no longer allow the identification of a sole ideal correction method, indicating changes in the underlying processes contributing to the bias. Our findings show that the correction approach yielding the lowest residual bias varies strongly across models and spatial domain. For example, while the QM method performs best for the CESM2-WACCM6 in the EU domain (Fig. 4b) the RB method yields a smaller residual bias in the US (Fig. 4f).

These results are supported by the analysis of the PDFs of the bias corrected model output (Fig. S8). While the conformity with observations remains widely similar for the majority of the distribution, the adjustment of the high tail yields slightly better results in context of the MB and RB methods, when compared to the base period. Contrarily, the distributions of both the DC and QM methods show a good agreement with the low tail and the midsection of the observational PDF. The performance however, is deteriorating towards the high tail, partially resulting in an overestimation of the monitored distribution, especially in the European domain.

To further investigate this curious result we examine, on a quantile basis across the MDA8 O_3 distributions, i) the error resulting from the initial bias correction of the base period (E_B) and ii) the error resulting from the deviation of the models change between the base and evaluation period when compared to observations (E_Δ).

$$E_B = M_q'^{hist} - O_q^{hist} \qquad (15)$$

$$E_F = M_q'^{proj} - O_q^{proj} \qquad (16)$$

$$= \left(M_q'^{hist} + \Delta M_q\right) - \left(O_q^{hist} + \Delta O_q\right)$$
$$= E_B + \Delta M_q - \Delta O_q = E_B + E_\Delta$$

The results of this analysis are exemplarily shown for the CESM2-WACCM6 model in Fig. 5 for the EU (a-c) and the US (d-f) (note the illustrations for the other models are included in the supplemental Fig. S9-10). Here the red shading indicates the minimum to maximum range of the residual bias across grid cells for the base period after bias correction (E_B, Eq. (15)), and the solid red line shows the domain average of this bias at individual quantiles concerned. In contrast the grey shading illustrates the minimum to maximum range of the differences in the change between the base and evaluation period of the raw model and observations, respectively (E_Δ, Eq. (16)). The black solid line marks the domain average of this bias at individual quantiles concerned. The residual bias for the evaluation period E_F (or in analogy any other future time period) comprises the sum of these errors (base bias and response bias) and is illustrated for the domain average with the dashed yellow line. We note that, as the DC method by definition yields no initial error in the base period only E_Δ is relevant in the evaluation period which is illustrated by the grey shading and the black solid line in all panels of Fig. 5 (as well as Fig. S9-10).

For the base period it is apparent that the QM correction technique, in contrast to RB and MB correction, yields only minor differences across the MDA8 O_3 distribution when compared to the observations in both spatial domains. For the evaluation period we see that the difference in response between models and observations is dominating over the raw performance of the individual correction techniques and that the residual bias depends strongly on region and model concerned (see Fig. 4-5 and supplemental Fig. S9-10). This illustrates the dependence of correction performance on models being able to represent precursor emission changes over time as seen in observations.

All models show distinct biases in reproducing observed ozone changes between the two time periods, in magnitude particularly pronounced in the tails of the distributions. Although both error terms, and the resulting net error are found to be rather small in the domain average (roughly ±5 ppb), they might have a strong influence on the individual grid cell level (see shadings). Especially for the MB and RB techniques the individual errors might compensate each other, as illustrated by the improved results relative to the base period. The DC and QM approaches on the other hand strongly depend on the quality of the model response in time. Here we find, that pronounced errors in the model change offset (at least in parts) the benefits illustrated for the base period (see Fig. 4).

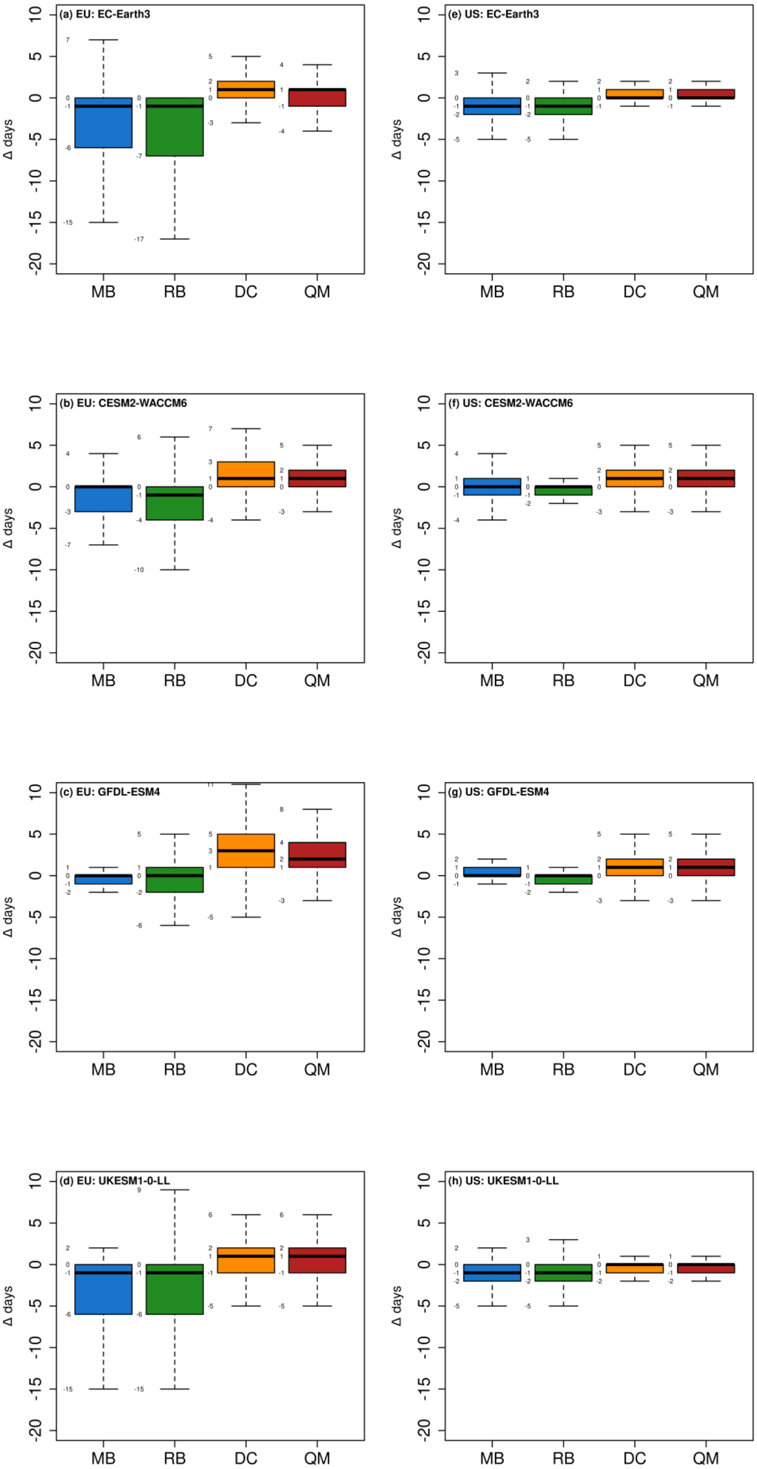

(a) EU: EC-Earth3
(e) US: EC-Earth3
(b) EU: CESM2-WACCM6
(f) US: CESM2-WACCM6
(c) EU: GFDL-ESM4
(g) US: GFDL-ESM4
(d) EU: UKESM1-0-LL
(h) US: UKESM1-0-LL
Δ days
10
5
0
-5
-10
-15
-20
MB
RB
DC
QM

Fig 4: As Fig. 3 but for 2010-2014 time period. Blue, Green, orange and red colour indicates the MB, RB, DC and QM correction methods respectively.

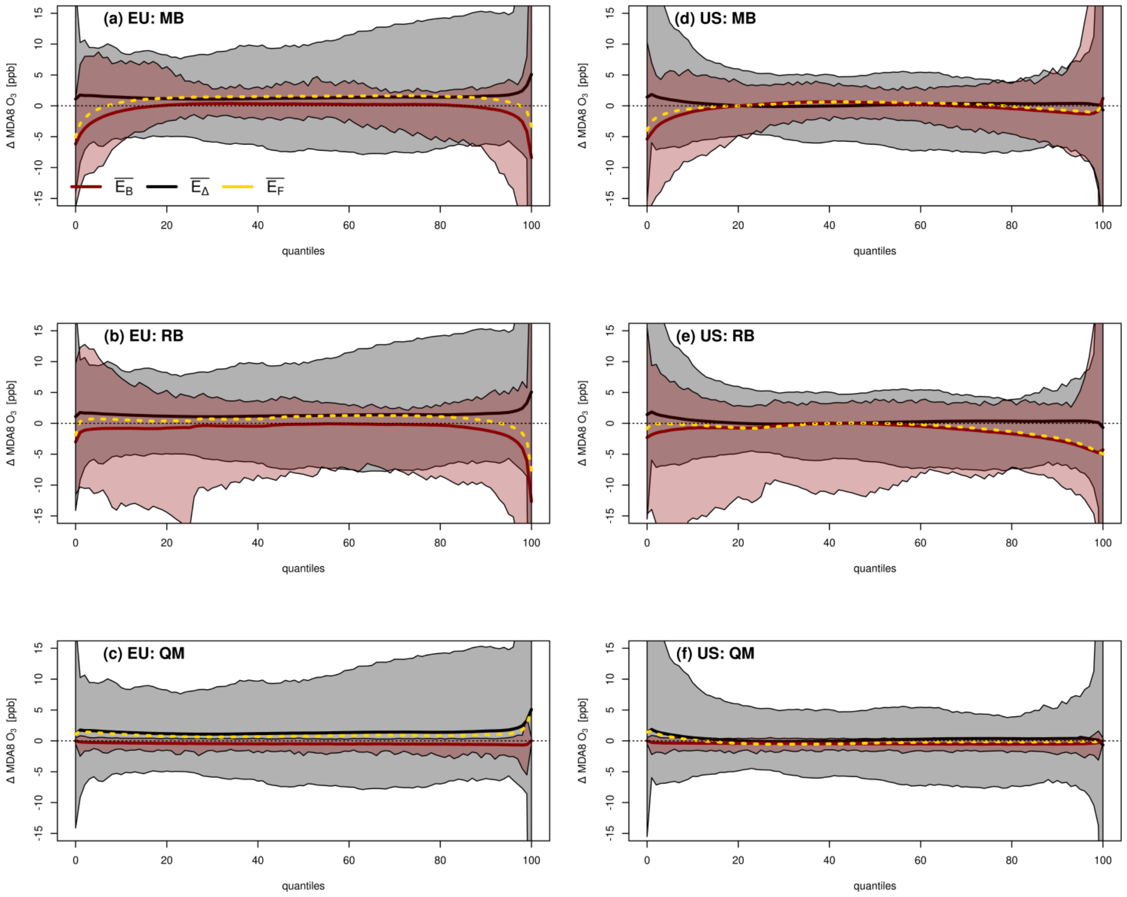


Fig. 5: Error Components of the CESM2-WACCM6 model during the evaluation period, for the MB (a & d), RB (b & e) and QM (c & f) methods, in the EU (left column) and US (right column) domain. The red shading gives the minimum to maximum range, and the solid red line the domain average of the residual bias in the base period (E_B), respectively. The grey shading gives the minimum to maximum range and the solid black line the domain average of the differences in the change between the base and evaluation period of the raw model and observations (E_Δ), respectively. The resulting domain average error of the evaluation period (E_F) is indicated by the dashed yellow line (note, for the DC method $E_B = 0$ and hence $E_F = E_\Delta$).

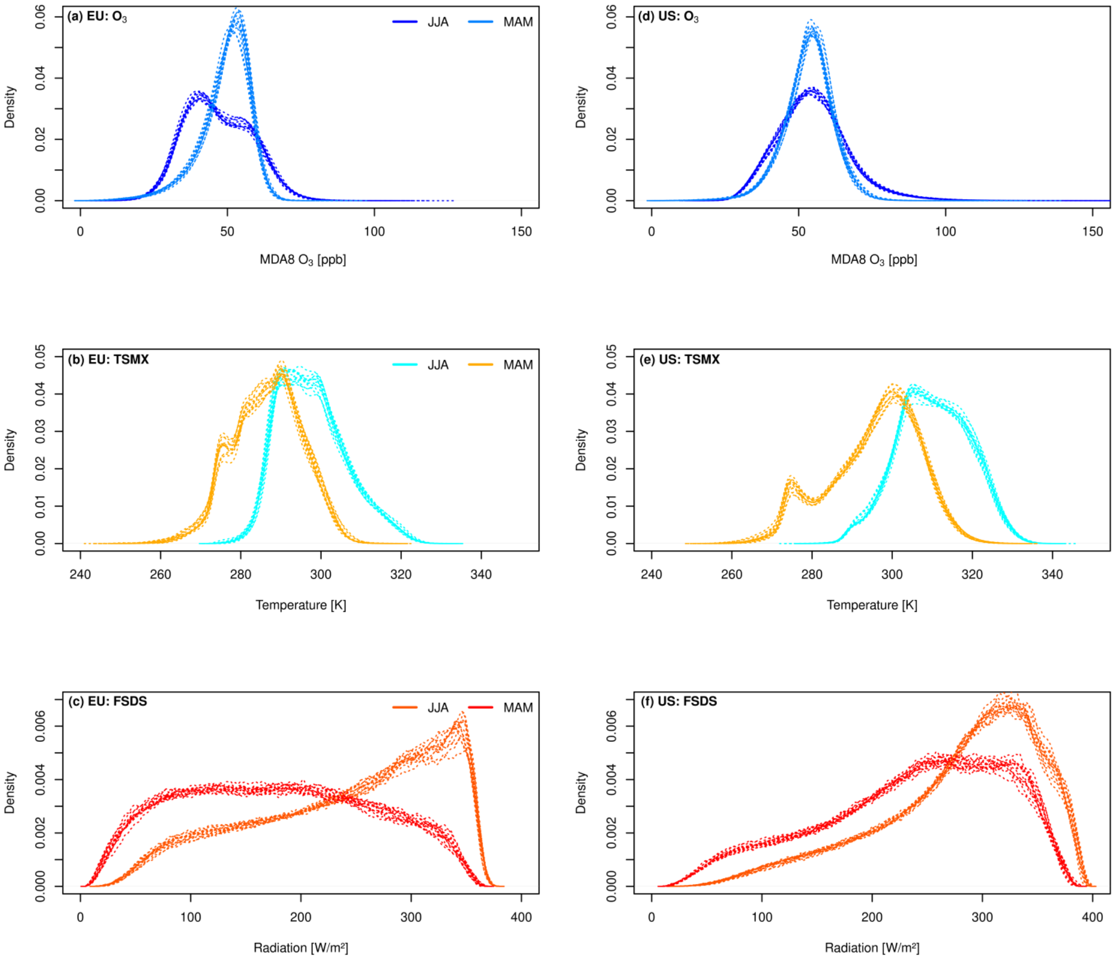
3.4. The influence of meteorology on the bias in the CESM2-WACCM6 ensemble

Having illustrated the MDA8 O_3 biases of various CMIP6 models, the performance of various statistical bias techniques as well as the influence of the model response to changes in e.g. emissions on the performance of bias correction we turn here to shed light on the underlying cause of biased MDA8 O_3 model outputs. To this end we analyse the 13 members of the CESM2-WACCM6 ensemble in more detail, in order to examine for consistency within the individual realizations as well as a possible dominant cause(s) for the bias in the modelled surface ozone fields. Here two likely prime candidates exist: 1) issues with the sensitivity in chemical mechanisms to local/regional precursor emissions, 2) issues in meteorological driving in the free running CCM. For the latter, we further include two climatological key drivers for ozone production in our analysis, i.e. daily maximum temperature (TSMX) and daily average down welling short wave radiation (FSDS), in order to differentiate whether the bias is predominantly driven by sensitivity to meteorology or chemistry.

Figure 6 illustrates the PDFs of MDA8 O_3, TSMX and FSDS for the individual ensemble members during spring and summer in 2005-2009 (the PDFs for 2010-2014 are shown in the supplemental Fig. S11). MAM and JJA MDA8 O_3 (Fig. 6a,d) show a very similar distribution across ensemble members for both domains. For example the median MDA8 O_3 value ranges across ensemble members roughly between 50 and 52 ppb (MAM) and 45 to 47 ppb (JJA) in the EU. For the US the median MDA8 O_3 values are found to be slightly higher than in the EU, but the differences within the ensemble lie in the same narrow range (53 to 55 ppb for MAM and 54 to 55 ppb for JJA). In context of the meteorological variables similar results are found. Although, slight variations occur for surface temperature and radiation (which one would expect from a model with free running meteorology), the PDFs are widely homogenous across the ensemble, thereby explaining as all ensemble members are driven with the same set of precursor emissions, the similarity of surface ozone distributions within the ensemble in both domains. The analysis of the MDA8 O_3, TSMX and FSDS distributions over the second time period (2010-2014, Fig. S11) yields similar results, thereby providing confidence for the robustness of our findings.

The strong similarity across ensemble members indicates that the MDA8 O_3 bias identified in CESM2-WACCM6 stems most likely from sensitivities in the chemical mechanism and/or emissions and not from meteorological drivers and their variability. Previous research has shown that temperature biases are rather small and that a significant overestimation of the temperature in the troposphere solely occurs in the southern hemisphere polar region [Danabasoglu et al., 2020; Gettelman et al., 2019], a region which is not investigated here. While the presented ensemble analysis is, due to data availability, only possible for CESM2-WACCM6, the results provide a first order estimate of the dominant model component responsible for surface ozone biases. Future work should confirm that this finding holds for other global models and thus an ensemble strategy for model experiments is recommended for future model intercomparison activities such as CCMI and CMIP.

https://doi.org/10.5194/egusphere-2023-2743
Preprint. Discussion started: 23 November 2023



335 **Fig. 6:** CESM2-WACCM6 spring (MAM) and summer time (JJA) PDFs of MDA8 O_3, TSMX and FSDS for the European (left) and US (right) domain in 2005-2009.

4. Summary & Conclusions

In this study we evaluate four global CCMs contributing to CMIP6 (EC-Earth3, CESM2-WACCM6, GFDL-ESM4 and UKESM1-0-LL) regarding their bias in surface ozone burdens, and present the first comprehensive comparison of the 340 performance of four different statistical bias correction techniques to derive CCM-based ozone metrics with relevance for public health and policy. While all models show biases when compared to observations, the bias magnitude of the raw, uncorrected MDA8 O_3 outputs differs strongly within the pool of models analyzed.

The evaluation of the four bias correction techniques for the base period (2005-2009), where techniques are tuned to observations, illustrates that all methods are capable to lower the bias. The MB and RB methods, however, are less accurate 345 when contrasted with the results obtained with the DC or QM approaches. Furthermore, when applying the MB and RB methods the model output fields might even be overcorrected for individual grid cells, i.e. the resulting ozone distributions might become biased low. This is not astonishing, as both techniques apply a single average value for the correction of the whole distribution function, which is a disadvantage - especially when it comes to the tails of the distribution - if the bias is not constant across the ECDF.

The independent evaluation of the four techniques over the second time period (2010-2014), focusing on the bias correction of model projections, yields less distinct results. Although the model to observation agreement is improved for all MDA8 O_3 metrics in the corrected models compared to their raw counterparts, no single optimal correction technique can be identified. Our results illustrate that technique performance depends strongly on model selected and its MDA8 O_3 evolution, and thus response to boundary conditions changes, over time. This at first surprising result, however, can be explained by the examination of the composition of the residual model error.

The residual error for future projections is comprised of two parts: 1) the residual error of the base period E_B, and 2) the error attributable to the model response to changes in boundary conditions (emissions, climate, etc…) between both time periods E_Δ. Especially the magnitude of E_Δ was found to exert a dominant influence on the overall correction performance, which raises some concerns regarding the robustness of model responses and thus the reliability of model projections (not only in context of surface O_3). In contrast to E_Δ, $E_B$depends on the quality of the initial base period bias correction. Here our results clearly show that $E_B$is substantially larger for the MB and RB than the QM and DC methods. When applying the correction techniques E_Δ and E_B might compensate for individual grid cells, resulting in a low residual bias. On the contrary, the strong base period performance obtained with the QM and DC approaches are attributable to a very low E_B, which might deteriorate in projections if E_Δ is large. Thus we conclude, that under the assumption of an adequate model response to changing boundary conditions (and thus low E_Δ), the QM and DC methods are outperforming the MB and RB techniques. If a decision has to be made whether the DC or QM approach is used for bias correction, we would argue, given that differences between the results obtained with both techniques are negligible, for DC correction due the comparably easy numerical implementation.

To obtain further insights of the root cause(s) of the bias surface ozone in models, we explored the MDA8 O_3 output of the 13 member CESM2-WACCM6 ensemble together with the key meteorological covariates for ozone production, i.e. daily maximum temperature and incoming shortwave radiation. Here our analysis showed only small variations within the CESM2-WACCM6 ensemble for core meteorological drivers of surface ozone. This suggests rather a dominant influence of the chemical mechanism and/or considered emissions on the bias in the O_3 fields than a prominent role for model meteorology. To investigate if this finding can be generalized to other CCMs is planned for future work as it also requires community efforts in providing additional ensemble simulations for individual CCMs contributing to the CCMI or CMIP frameworks.

Data availability

CMIP6 data sets are publicly available via https://esgf-data.dkrz.de/projects/cmip6-dkrz/. Processed data can be made available by the corresponding author upon reasonable request.

Author Contribution

C. Staehle: conceptualization, formal analysis, methodology, visualization, writing – original draft preparation

H. E. Rieder: conceptualization, methodology, resources, supervision, writing – review and editing

A. M. Fiore: resources, supervision, writing – review and editing

Competing Interests

The authors declare that they have no conflict of interest.

Acknowledgement

The authors are grateful to the EC-Earth3, GFDL-ESM4 and UKESM1-0-LL modelling teams for providing the ensemble simulations via https://esgf-data.dkrz.de/projects/cmip6-dkrz/ and to J. Schnell for providing the gridded surface ozone data set. H.E. Rieder and C. Staehle acknowledge support by the Austrian Climate and Energy Fund via project ACRP11-KR18AC0K14686. C. Staehle acknowledges support through an OeAD Marietta Blau Fellowship grant. The authors are grateful to Ramiro Checa-Garcia for fruitful discussions and comments.

References

Archibald, A. T., Neu, J. L., Elshorbany, Y. F., Cooper, O. R., Young, P. J., Akiyoshi, H., Cox, R. A., Coyle, M., Derwent, R. G., Deushi, M., Finco, A., Frost, G. J., Galbally, I. E., Gerosa, G., Granier, C., Griffiths, P. T., Hossaini, R., Hu, L., Jöckel, P., Josse, B., Lin, M. Y., Mertens, M., Morgenstern, O., Naja, M., Naik, V., Oltmans, S., Plummer, D. A., Revell, L. E., Saiz-Lopez, A., Saxena, P., Shin, Y. M., Shahid, I., Shallcross, D., Tilmes, S., Trickl, T., Wallington, T. J., Wang, T., Worden, H. M., Zeng, G. (2020). Tropospheric Ozone Assessment Report: A critical review of changes in the tropospheric ozone burden and budget from 1850 to 2100. *Elementa: Science of the Anthropocene, 8*(1), 034. doi:10.1525/elementa.2020.034

Chameides, W. L., Lindsay, R. W., Richardson, J., Kiang, C. S. (1988). The Role of Biogenic Hydrocarbons in Urban Photochemical Smog: Atlanta as a Case Study. *Science, 241*(4872), 1473-1475. doi:10.1126/science.3420404

Checa-Garcia, R., Hegglin, M. I., Kinnison, D., Plummer, D. A., Shine, K. P. (2018). Historical Tropospheric and Stratospheric Ozone Radiative Forcing Using the CMIP6 Database. *Geophys Res Lett, 45*(7), 3264-3273. doi:10.1002/2017GL076770

Da, Y., Xu, Y., McCarl, B. (2022). Effects of Surface Ozone and Climate on Historical (1980–2015) Crop Yields in the United States: Implication for Mid-21st Century Projection. *Environmental and Resource Economics, 81*(2), 355-378. doi:10.1007/s10640-021-00629-y

Danabasoglu, G., Lamarque, J. F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., Emmons, L. K., Fasullo, J., Garcia, R., Gettelman, A., Hannay, C., Holland, M. M., Large, W. G., Lauritzen, P. H., Lawrence, D. M., Lenaerts, J. T. M., Lindsay, K., Lipscomb, W. H., Mills, M. J., Neale, R., Oleson, K. W., Otto-Bliesner, B., Phillips, A. S., Sacks, W., Tilmes, S., van Kampenhout, L., Vertenstein, M., Bertini, A., Dennis, J., Deser, C., Fischer, C., Fox-Kemper, B., Kay, J. E., Kinnison, D., Kushner, P. J., Larson, V. E., Long, M. C., Mickelson, S., Moore, J. K., Nienhouse, E., Polvani, L., Rasch, P. J., Strand, W. G. (2020). The Community Earth System Model Version 2 (CESM2). *J Adv Model Earth Sy, 12*(2), e2019MS001916. doi:10.1029/2019MS001916

EEA. (2020). *Air quality in Europe - 2020 Report* (09/2020). Retrieved from https://www.eea.europa.eu/publications/air-quality-in-europe-2020-report

EPA. (2015). *National Ambient Air Quality Standards for Ozone*. Retrieved from https://www.govinfo.gov/content/pkg/FR-2015-10-26/pdf/2015-26594.pdf

EUR-LEX. (2011). *Commission Implementing Decision of 12 December 2011 laying down rules for Directives 2004/107/EC and 2008/50/EC of the European Parliament and of the Council as regards the reciprocal exchange of information and reporting on ambient air quality (notified under document C(2011) 9068)* (2011/850/EU). Retrieved from http://data.europa.eu/eli/dec_impl/2011/850/2011-12-17

Fleming, Z. L., Doherty, R. M., von Schneidemesser, E., Malley, C. S., Cooper, O. R., Pinto, J. P., Colette, A., Xu, X., Simpson, D., Schultz, M. G., Lefohn, A. S., Hamad, S., Moolla, R., Solberg, S., Feng, Z. (2018). Tropospheric Ozone Assessment Report: Present-day ozone distribution and trends relevant to human health. *Elementa: Science of the Anthropocene, 6*, 12. doi:10.1525/elementa.273

Gettelman, A., Mills, M. J., Kinnison, D. E., Garcia, R. R., Smith, A. K., Marsh, D. R., Tilmes, S., Vitt, F., Bardeen, C. G., McInerny, J., Liu, H. L., Solomon, S. C., Polvani, L. M., Emmons, L. K., Lamarque, J. F., Richter, J. H., Glanville, A. S., Bacmeister, J. T., Phillips, A. S., Neale, R. B., Simpson, I. R., DuVivier, A. K., Hodzic, A., Randel, W. J.

(2019). The Whole Atmosphere Community Climate Model Version 6 (WACCM6). *J Geophys Res-Atmos, 124*(23), 12380-12403. doi:10.1029/2019JD030943

Griffiths, P. T., Murray, L. T., Zeng, G., Shin, Y. M., Abraham, N. L., Archibald, A. T., Deushi, M., Emmons, L. K., Galbally, I. E., Hassler, B., Horowitz, L. W., Keeble, J., Liu, J., Moeini, O., Naik, V., O'Connor, F. M., Oshima, N., Tarasick, D., Tilmes, S., Turnock, S. T., Wild, O., Young, P. J., Zanis, P. (2021). Tropospheric ozone in CMIP6 simulations. *Atmos Chem Phys, 21*(5), 4187-4218. doi:10.5194/acp-21-4187-2021

Lehner, F., Nadeem, I., Formayer, H. (2023). Evaluating skills and issues of quantile-based bias adjustment for climate change scenarios. *Adv Stat Clim Meteorol Oceanogr, 9*(1), 29-44. doi:10.5194/ascmo-9-29-2023

Lelieveld, J., Dentener, F. J. (2000). What controls tropospheric ozone? *J Geophys Res-Atmos, 105*(D3), 3531-3551. doi:10.1029/1999JD901011

Liu, Z., Doherty, R. M., Wild, O., O'Connor, F. M., Turnock, S. T. (2022). Correcting ozone biases in a global chemistry–climate model: implications for future ozone. *Atmos Chem Phys, 22*(18), 12543-12557. doi:10.5194/acp-22-12543-2022

Meehl, G. A., Tebaldi, C., Tilmes, S., Lamarque, J.-F., Bates, S., Pendergrass, A., Lombardozzi, D. (2018). Future heat waves and surface ozone. *Environ Res Lett, 13*(6), 064004. doi:10.1088/1748-9326/aabcdc

Mills, G., Pleijel, H., Malley, C. S., Sinha, B., Cooper, O. R., Schultz, M. G., Neufeld, H. S., Simpson, D., Sharps, K., Feng, Z., Gerosa, G., Harmens, H., Kobayashi, K., Saxena, P., Paoletti, E., Sinha, V., Xu, X. (2018). Tropospheric Ozone Assessment Report: Present-day tropospheric ozone distribution and trends relevant to vegetation. *Elementa: Science of the Anthropocene, 6*, 47. doi:10.1525/elementa.302

Monks, P. S., Archibald, A. T., Colette, A., Cooper, O., Coyle, M., Derwent, R., Fowler, D., Granier, C., Law, K. S., Mills, G. E., Stevenson, D. S., Tarasova, O., Thouret, V., von Schneidemesser, E., Sommariva, R., Wild, O., Williams, M. L. (2015). Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer. *Atmos Chem Phys, 15*(15), 8889-8973. doi:10.5194/acp-15-8889-2015

Rieder, H. E., Fiore, A. M., Clifton, O. E., Correa, G., Horowitz, L. W., Naik, V. (2018). Combining model projections with site-level observations to estimate changes in distributions and seasonality of ozone in surface air over the U.S.A. *Atmos Environ, 193*, 302-315. doi:10.1016/j.atmosenv.2018.07.042

Rieder, H. E., Fiore, A. M., Horowitz, L. W., Naik, V. (2015). 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. *J Geophys Res-Atmos, 120*(2), 784-800. doi:10.1002/2014JD022303

Schnell, J. L., Holmes, C. D., Jangam, A., Prather, M. J. (2014). Skill in forecasting extreme ozone pollution episodes with a global atmospheric chemistry model. *Atmos Chem Phys, 14*(15), 7721-7739. doi:10.5194/acp-14-7721-2014

Sillman, S. (1999). The relation between ozone, NOx and hydrocarbons in urban and polluted rural environments. *Atmos Environ, 33*(12), 1821-1845. doi:10.1016/S1352-2310(98)00345-8

Sillman, S., Logan, J. A., Wofsy, S. C. (1990). The sensitivity of ozone to nitrogen oxides and hydrocarbons in regional ozone episodes. *J Geophy Res-Atmos, 95*(D2), 1837-1851. doi:10.1029/JD095iD02p01837

Turnock, S. T., Allen, R. J., Andrews, M., Bauer, S. E., Deushi, M., Emmons, L., Good, P., Horowitz, L., John, J. G., Michou, M., Nabat, P., Naik, V., Neubauer, D., O'Connor, F. M., Olivié, D., Oshima, N., Schulz, M., Sellar, A., Shim, S., Takemura, T., Tilmes, S., Tsigaridis, K., Wu, T., Zhang, J. (2020). Historical and future changes in air pollutants from CMIP6 models. *Atmos Chem Phys, 20*(23), 14547-14579. doi:10.5194/acp-20-14547-2020

Westervelt, D. M., Ma, C. T., He, M. Z., Fiore, A. M., Kinney, P. L., Kioumourtzoglou, M. A., Wang, S., Xing, J., Ding, D., Correa, G. (2019). Mid-21st century ozone air quality and health burden in China under emissions scenarios and climate change. *Environ Res Lett, 14*(7), 074030. doi:10.1088/1748-9326/ab260b

Young, P. J., Naik, V., Fiore, A. M., Gaudel, A., Guo, J., Lin, M. Y., Neu, J. L., Parrish, D. D., Rieder, H. E., Schnell, J. L., Tilmes, S., Wild, O., Zhang, L., Ziemke, J., Brandt, J., Delcloo, A., Doherty, R. M., Geels, C., Hegglin, M. I., Hu, L., Im, U., Kumar, R., Luhar, A., Murray, L., Plummer, D., Rodriguez, J., Saiz-Lopez, A., Schultz, M. G., Woodhouse, M. T., Zeng, G. (2018). Tropospheric Ozone Assessment Report: Assessment of global-scale model performance for global and regional ozone distributions, variability, and trends. *Elementa: Science of the Anthropocene, 6*, 10. doi:10.1525/elementa.265