General comments:

We thank the reviewers for their constructive comments and have revised the manuscript in response. Our improved manuscript includes the addition of a section dedicated to understanding the model in more detail. We also thank the reviewers for their time and patience during the process.

In this document, we first address a major point from both reviewers to include more process understanding, then we respond to reviewers comments sequentially. Process understanding is achieved by further analysis exploring relationships between mean ozone, the DOR and NOx concentrations, in addition to comparison of satellite NO2 columns. This is included at the end of Sect. 3.4 (lines 340-380) and inserted below:

"To examine possible reasons for (i) the bias in the monthly means and (ii) the worse performance of the DOR at some sites, we consider how these variables are related to NOx concentrations. At the majority of remote sites in the tropics, ozone production is controlled by NOx concentrations i.e., with the exception of Watukosek, the sites are NOx-limited because ozone production rate increases with increasing NOx (Fig. 7a). Here, ozone production rate is defined as the rate of reaction NO + RO_2/HO_2 with NO controlling seasonality at NOx-limited sites. At Watukosek, the seasonality in ozone production rate is less clearly attributable to NOx concentrations, which indicates other factors (such as VOC concentration or meteorology) are involved.

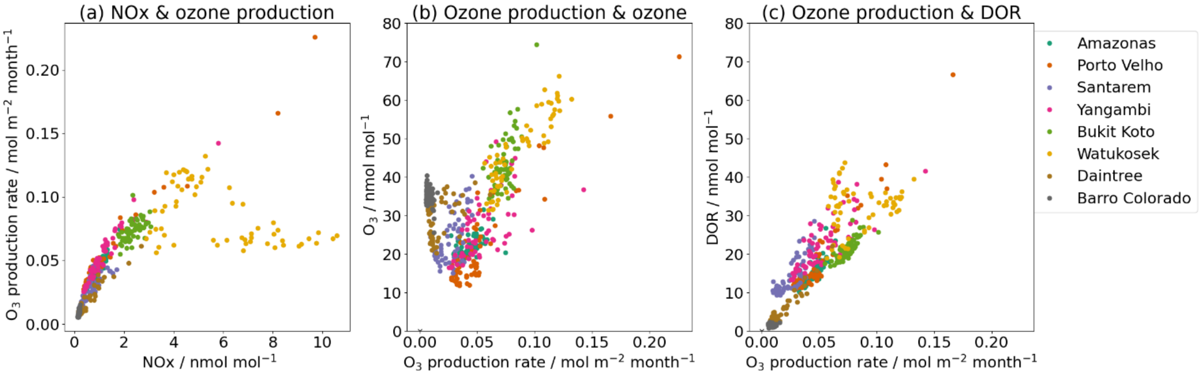


Fig. 7: Relationship between monthly mean ozone production rate at the surface at each remote site for (a) surface NOx concentration, (b) surface ozone concentration and (c) the diurnal ozone range (DOR).

However, seasonal patterns in mean ozone concentrations can differ from ozone production rate due to changes in chemical loss rate and non-chemical factors such as deposition and transport. Two sites that highlight this in Fig. 7b are Daintree and Barro Colorado; the low ozone production rates suggest that transport to these sites cause ozone concentrations to be high and unrelated to seasonality in ozone production. As coastal sites, they are likely to be strongly influenced by coastal weather phenomena, which may include thermally-driven transport. However, for these sites the DOR, which shows different seasonality to mean ozone concentrations (c.f. Fig. 5 and Fig. 6), is correlated with the seasonal changes in ozone production (Fig. 7c). This suggests that the DOR can be useful in understanding local-scale processes. In particular, NOx concentration is likely a major factor affecting the DOR. In fact, even at Watukosek, the seasonal cycle in the DOR is correlated with NOx concentrations (r^2 = 0.58; Fig. S8). This perhaps suggests that NO availability and its change with daily insolation, is the main factor affecting the seasonal cycle in the DOR. Of course, loss processes *must also play a role in the DOR magnitude, however it certainly seems that, at these rural sites, ozone production and the DOR have strong relationships to seasonal NOx concentration.*

Therefore, sites with worse performance by UKESM1 at reproducing the DOR may indicate poor representation of local NOx chemistry. These sites include Yangambi, Watukosek and Bukit Koto, so we compare their tropospheric NO_2 columns in UKESM1 to OMI satellite products (Fig. S9). Yangambi and Bukit Koto show different seasonality in the NO_2 columns at the site compared to the satellite product. This indicates there could be an issue with prescribed emissions of NOx, or that NOx processes are poorly represented at these sites. Additionally, we find that NO_2 columns from UKESM1 are approximately 3x higher than the satellite columns, which may signify that NOx concentrations are too high in the model, although not necessarily at the surface. Inefficient boundary layer mixing of surface emitted species may contribute to the aggregation of NOx in near-surface model levels. Given the relationship between NOx and ozone production (Fig. 7a), a decrease in NOx concentration would likely result in a decrease in ozone concentrations. However, it may also decrease the DOR and therefore would not be the only cause of the differences between modelled and observed ozone. It is also worth noting that NO_2 columns are sensitive to the algorithm used to calculate the columns, including the method used to separate tropospheric from stratospheric NO_2. Although a systematic high background NO_2 in the troposphere may be a cause of the systematic ozone bias, more observations of NOx are needed to confirm this, as well as to understand whether a bias is related to emissions, the physical model or chemistry. Further discussion in relation to a positive NOx bias from the literature is included in Sect. 4.3. On the other hand, the poor model performance of ozone seasonality and the DOR at Yangambi and Watukosek does seem likely to contribute to stem from incorrect representation of the NOx seasonality by UKESM1. In four other Earth system models performing the same simulation, the seasonality at Watukosek is captured better, whereas all models overestimate the change in ozone during biomass burning months at Yangambi, with UKESM1 performing among the best on account of the smaller seasonal variation (Fig. S10)."

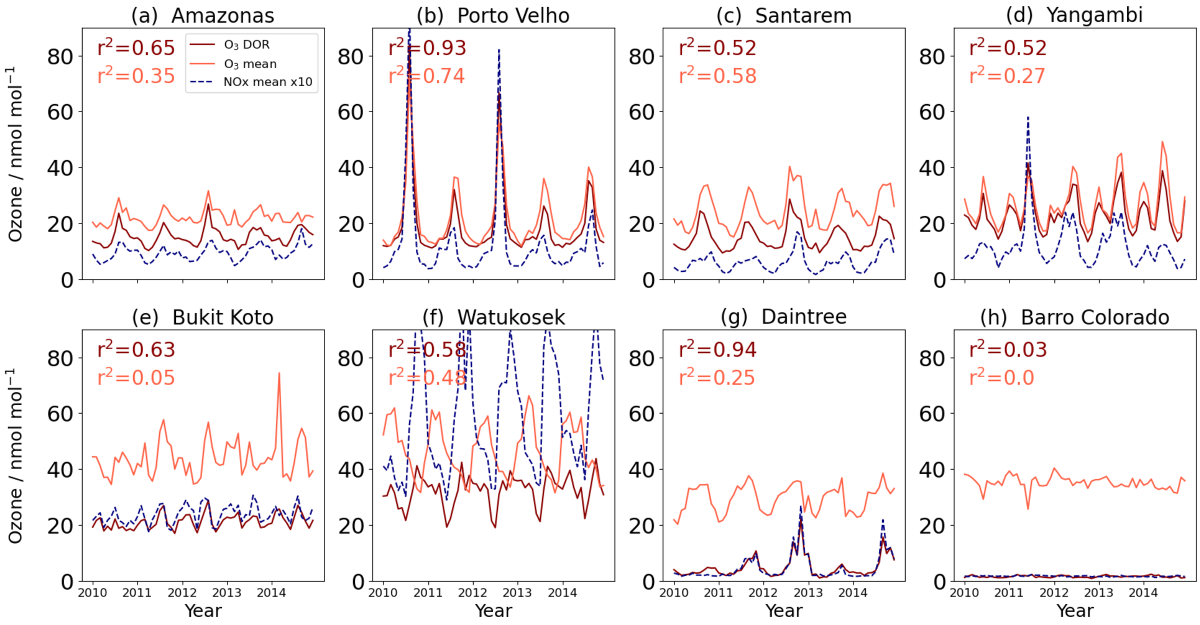


Figure S8: Monthly mean data between 2010 – 2014 showing patterns in ozone concentrations (orange solid line), the DOR (red solid line) and NOx concentrations (x10, blue dashed line). R^2 values in the upper left corner gives the correlation between NOx and the DOR (red) and ozone (orange).

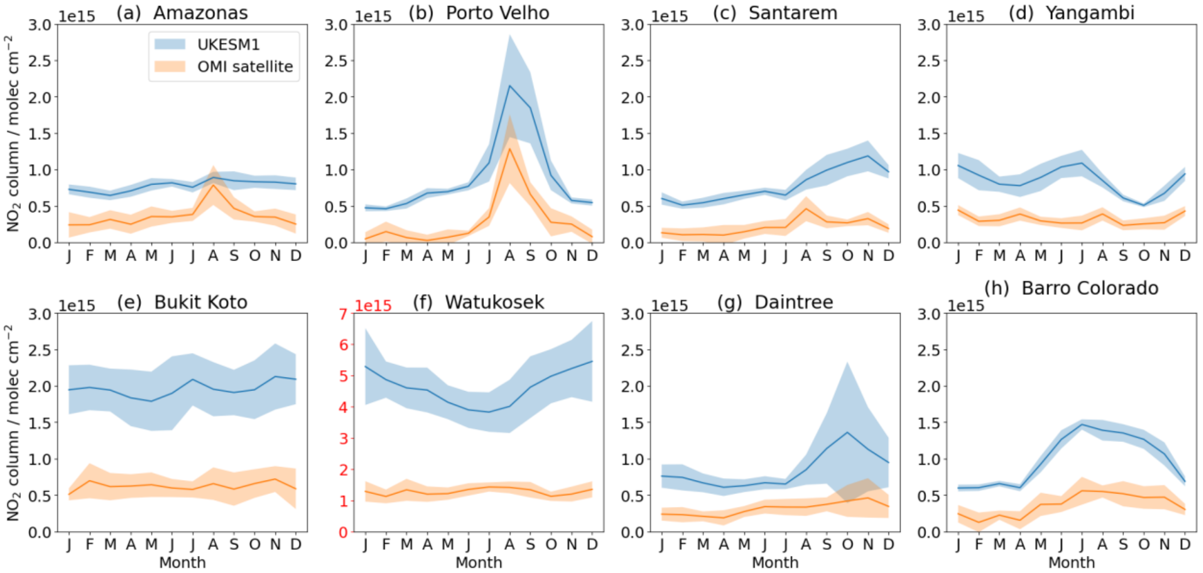


Figure S9: Monthly mean tropospheric NO₂ columns between 2004 – 2014 at each site for UKESM1 (blue solid line) and OMI satellite (orange solid line). Shading covers 1 standard deviation using monthly means. Note the different scale used for Watukosek to improve readability.

Reviewer 1:

This manuscript describes an assessment of surface ozone in the Tropics from the UKESM1 model, demonstrating that mixing ratios are systematically high across all sites but that the average diurnal variation is reproduced well. It is a competent study, and scientifically sound, but it provides little fresh insight or understanding of either model performance or atmospheric behavior. A few simple sensitivity studies to investigate the cause of the biases would have made the paper substantially stronger, and this is a missed opportunity. It is highly likely that monthly mean averaging of biomass burning emissions is one source of error, as speculated, and that representation of the vertical profile in the lower boundary layer is another, but no solid evidence is provided. I feel that this needs to be addressed to at least some extent before the paper is suitable for publication.

Thank you for your comment and advice to include more insight and model understanding in the paper. While we were unable to perform sensitivity studies using UKESM1, we looked in detail at related model variables and their relationship to O3 concentrations, including NOx and O3 production rate. This is included as an additional paragraph that has been inserted above and found in the test within Sect. 3.4 (lines 340-380). It is focused on the monthly timescale. The main conclusions are:

- NOx concentration is highly related to the rate of O3 production and the DOR, showing the area is NOx-limited. However, these variables are not always strongly related to the overall monthly mean O3 concentration since this is affected by non-local processes (i.e., transport).

- Therefore, the DOR captures local processes that vary at hourly timescales, which is slightly different to mean O3 concentrations and allows us to consider these processes separately. We feel this is therefore a useful metric in addition to O3 concentration that should be considered more frequently.
- Models that perform poorly at representing the seasonal cycle in the DOR appear to be capturing NO chemistry incorrectly after comparison with OMI tropospheric satellite NO2 columns.
- Tropospheric NO2 columns are 3x higher in UKESM1 compared to the OMI satellite retrieval. This may be related to boundary layer representation and model resolution that affects the ability of NOx to leave the lowest layer. It may be a source of model bias but there are many other possibilities, including deposition, missing chemistry and representation of convection and vertical transport within the model.
- Systematic bias is present even at sites with low O3 production, confirming a synoptic bias across the tropics that is not location specific. The source of bias must act over a large amount of the tropics, but not necessarily a result of missing processes at the locations evaluated.

At this time, studies to diagnose the cause of model bias do not seem feasible. The present work was conducted using freely available data from CMIP6 and therefore did not require access to the model. The simulations used in this study were designed and run by modelling centres, with input data and model parameters carefully chosen. To set up the model in the same way would be time consuming, and perhaps altogether impossible given I, the first author, no longer work in the UK. Although some possible causes of model bias such as reducing fire emissions would be slightly easier to implement, the majority of issues include code development such as adding new chemistry and altering deposition terms or redesigning inputs at different spatial and temporal resolutions. To contact the modelling centres, set up the model and to replicate these studies under varying conditions to test possible model bias is a large task when the causes of bias are as broad as they are. Nonetheless, we made substantial efforts to address the concerns raised by the reviewers using the resources available.

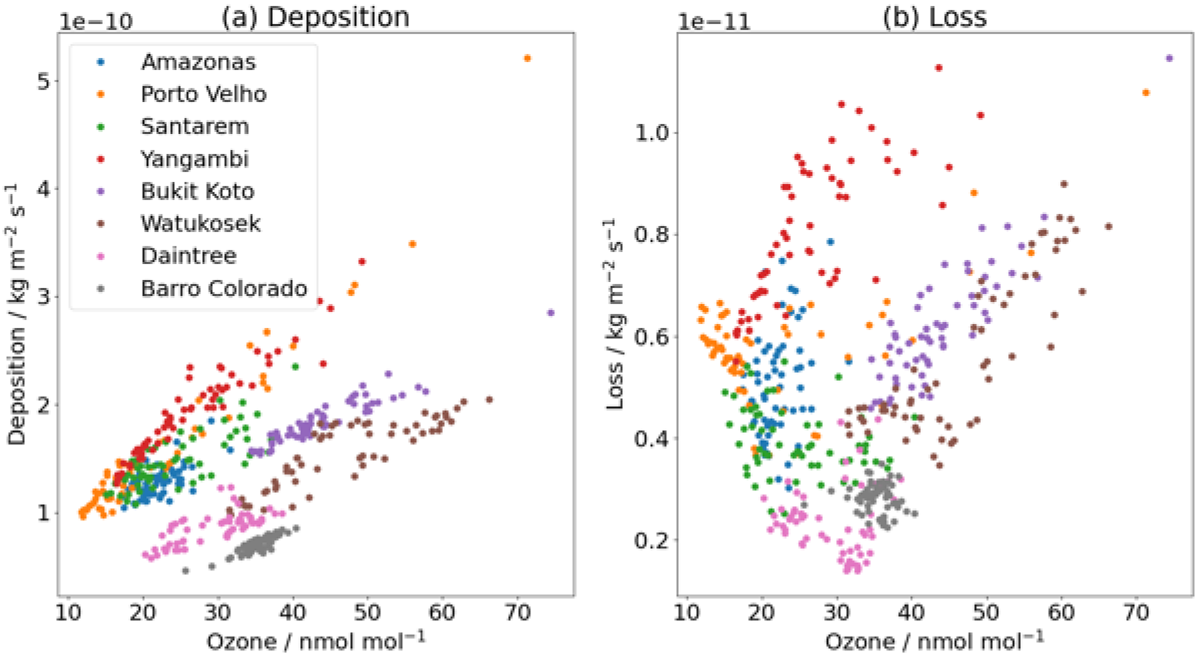
General Comments

One premise of the paper is that reasonable representation of the diurnal ozone range is a strength, despite a factor of two overestimation of ozone mixing ratios. Given that both deposition processes and titration by NO are first order in ozone, there is reason to expect proportional biases in DOR and ozone. Why should good representation of DOR give us confidence in the model, and what does this reveal about the source of the bias?

This is an interesting comment as it is true that many possible changes to the mean O3 may also change the DOR. It is indeed true that deposition is first order in O3, however overall chemical and deposition losses depend on several factors (see fig below in which chemical loss is unrelated to O3 concentration at a monthly timescale). Also, considering that the DOR represents how these processes change over the diurnal cycle and that processes such as deposition change in magnitude over the day, it is more difficult to predict how systematic bias would affect these relative processes. Together, this means that the DOR is probing more than just biases in the 1st-order chemistry and thus it goes significantly beyond analysing the bias in the mean state. There may not be a proportional bias in both the DOR and mean O3 at the same time if the DOR is less sensitive to, or independent of the mean state.

To address this further, we have added to Sect. 3.4 to suggest that model mean O3 is related to local and background processes whereas the DOR represents processes at a short time scale and thus usually local-scale processes. In some cases, the seasonal cycle in the DOR bears no resemblance to the mean O3, likely as a result of transport. Therefore, overestimation in the mean state could be a result of missing process representation elsewhere that increases background O3 through transport e.g. underestimating O3 deposition over the ocean. We also show that the mean bias in models is unrelated to their ability to represent seasonal variability (Fig. S9) and so we feel that the difference between variability and mean state through evaluation of the DOR is an important message within the paper.

Following your comments above, the high O3 concentrations may be reduced while maintaining the DOR by including an additional loss process. This is a feasible option and there are several ways to introduce this: increasing ocean deposition/chemistry as stated above, or missing in-canopy loss processes (e.g. sesquiterpene or SOA chemistry). In reality, there are likely several causes for the positive O3 bias and it is difficult to confirm the source of the bias from this information.



Fig(not in manuscript): Monthly mean variables at each site between 2005 – 2014 showing relationships between (a) ozone concentration and ozone dry deposition rate and (b) ozone concentration and ozone chemical loss rate.

This is now discussed on line 413:

"This study shows that analysis of the DOR provides unique information on the seasonality in NOx concentrations and ozone production at these remote tropical sites. Seasonality in the DOR is strongly related to NOx concentrations, demonstrating changes in NO concentration over the day is a significant contributor to the DOR. Overnight, absence of photolysis prevents the ozone-producing reaction NO + RO_2/HO_2 as NO is locked up in NO_2 and reservoir species, causing ozone concentrations to decline. During the day, NO is formed through NO_2 photolysis and the maximum rate of ozone production is determined partly by the NOx concentration, allowing a greater diurnal increase in months with higher NOx. This is by no means the only process controlling the DOR, but suggests that changes in local chemistry, such as NOx chemistry and subsequent ozone formation, seems to be captured by UKESM1.

Furthermore, the systematic ozone bias is present even in locations with low ozone production / where the ozone seasonal cycle is dominated by transport (e.g. Barro Colorado). Previous studies have indicated a bias at remote ocean sites of 10 ppb (Brown et al., 2022) indicating a background bias that likely extends across large parts of the tropics and does not necessarily originate at the site."

A case is made for more measurements of surface ozone across Tropical regions. While these would certainly be valuable, this need is not an outcome arising from this study, where it is clear that an otherwise well-tested model is unable to reproduce the surface ozone measurements we already have. A more relevant call would therefore be to investigate and diagnose these biases in models to fully explain and ideally eliminate it (and I believe that this paper should make the first steps towards doing this).

The need for more measurements may indeed be more subjective than the need to improve the model. We mention this as we feel that one of the limitations is understanding the homogeneity and variability within a gridcell, meaning we do not know how well the single observation point represents the gridcell. More measurements would therefore give information on the extent to which resolution is driving model bias. Additional measurements would also indicate whether the poor representation in Yangambi is local or regional. With a single observation station for Africa there is no way to know this. Of the observation data that are present, most have a limited time range, often including large gaps for maintenance, making comparison to model data more challenging and preventing adequate quantification of error and interannual variability. We feel more data would be useful in reducing some of the uncertainty in our study.

We highlight that in addition to being an evaluation of UKESM1 at several new sites, we also bring together a range of observations of tropical O3 concentrations that have not previously been brought together one publication. Consequently, this study aims to represent the tropical O3 data available, including by drawing attention to the limited availability.

That said, investigation of systematic bias is another good suggestion. The cause of the bias has been investigated for some time, with many possible options that we were unable to fulfil in one study. However, we do give some evidence to suggest that NOx build up in the surface layer may be partly responsible (see earlier comments and additional Sect. 3.4). Flagging these problems is the first and most important step in addressing model biases or poor process representation. Thus we hope that this manuscript, while not able to explain biases at every station, is helpful to model developers. The next step of understanding theses biases in detail is a long process that is beyond the scope of this paper, but is intended for future work.

It is clear that surface ozone from global models such as UKESM1 cannot be used for reliable health and ecosystem impact assessments at the current time. The paper manages to arrive at the opposite conclusion (line 457). While bias correction can remove model errors effectively, it can also remove the need for a model in the first place (there are now a wealth of machine-learned observation-based surface ozone climatologies available). A much stronger argument for use of models is needed, otherwise the point made here needs to be substantially reframed in the paper.

As described earlier, we have added extra text to show that among several models, the seasonality and ability to capture variability is not linked to the magnitude of the systematic bias. Whilst UKESM1 has a significant positive bias, it captures the seasonality at the sites fairly well. If systematic bias is accounted for, we can still use the model to understand variability and provide process understanding. We have adjusted the phrasing on line 578 to be clearer that the model cannot be used directly:

"Analysis of the DOR allows local-scale responses to be considered separately to the systematic bias and may be a useful diagnostic for other researchers to consider. Overall, our results suggest that UKESM1 may be useful for understanding ozone responses to forcings but hourly data should not be used 'off the shelf' for health and ecosystem impact assessments. Bias correction may be an option to avoid overestimation of the risks but users should be aware that monthly mean concentrations may require multiplicative bias correction in biomass burning regions and that gridcells containing non-homogeneous emission sources or land cover types may be impacted by the negative effects of coarse model resolution more than pristine regions. The magnitude of the bias in different regions and seasons, and its dependence on factors such as distance from emissions sources remains to be quantified. For this, more in situ monitoring is instrumental. "

Models provide process understanding that is simply not available using ML techniques, and typically perform better at projecting future changes, which is useful for policy. Furthermore, a model is still highly necessary in the tropics due to lack of measurement coverage and lack of ML products that perform well in the tropics. ML requires a lot of data to train and test a model, and where there are gaps (the whole of Africa, for example) developers often rely on reanalysis. Recently, several reanalysis products were shown to have no resemblance to O3 at the Yangambi site in the Congo (Vieira et al., 2023) because reanalysis also requires at least some observations. ML observation-based surface O3 is therefore not a feasible option in the tropics at this time, as far as we are aware.

Vieira, Inês, et al. "Global reanalysis products cannot reproduce seasonal and diurnal cycles of tropospheric ozone in the Congo Basin." *Atmospheric Environment* 304 (2023): 119773.

Which of the stations considered are expected to be representative of the model grid used here (based on geospatial homogeneity of land cover and emissions) and which ones are not? These issues are touched on in section 4.2, but no explicit assessment is made, and this makes interpretation of biases more difficult.

From analysis of GFED4s burned area within each gridcell, we identify that Yangambi and Porto Velho have a small burned area fraction likely indicating infrequent small fire activity, and a larger burned area in Daintree. This indicates a time-dependent and inhomogeneous emissions source at these sites. We also show from the land cover fractions produced by the model that some crop area is present in the S.E. Asian gridcells, indicating an inhomogeneous distribution of landcover. This figure also shows that local coastal effects may affect Bukit Koto, Watukosek, Daintree and Barro Colorado since the gridcells are coastal. This is not a comprehensive analysis, but we would recommend that sites at Santarem, Amazonas are the most homogeneous.

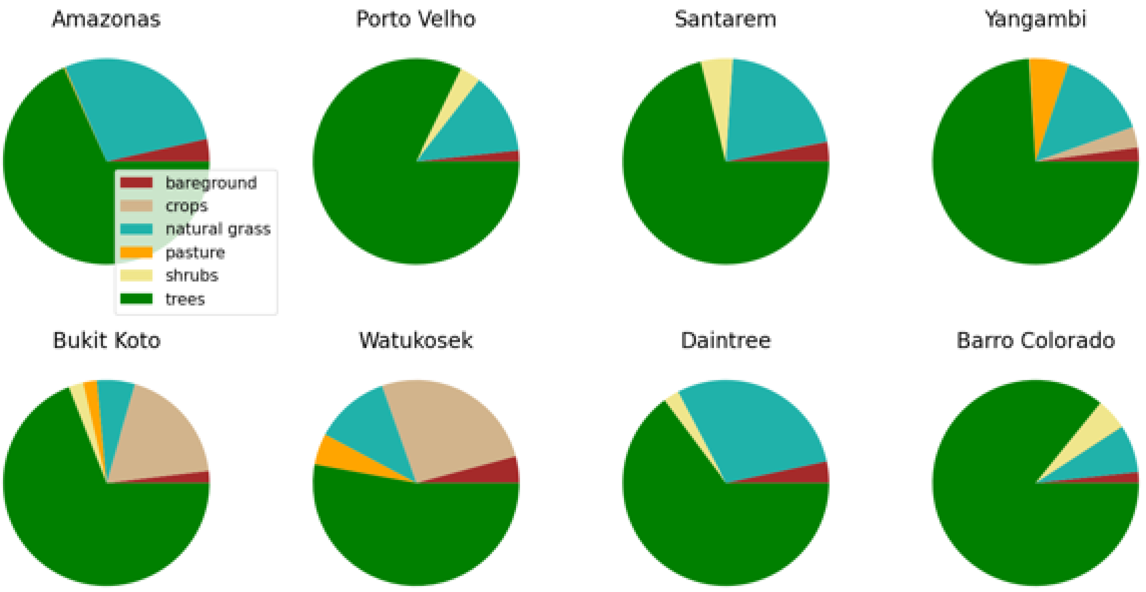


Fig (not included in manuscript): Percentage landcover in each UKESM1 gridcell

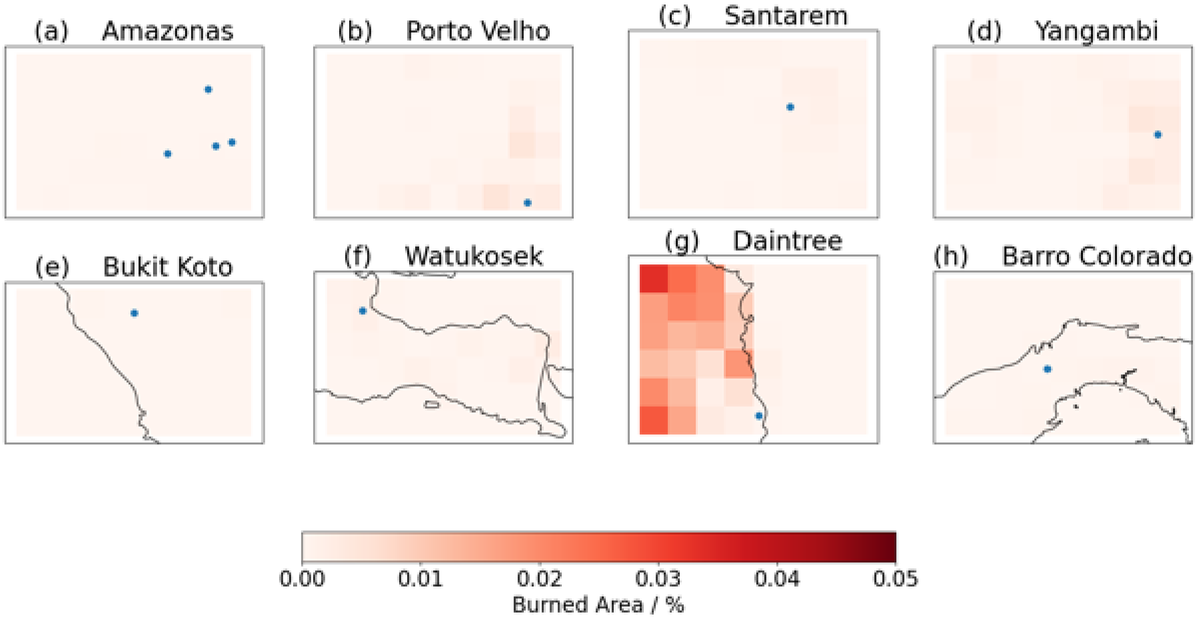


Fig (not included in manuscript): Mean percentage Burned Area from GFED4s within each gridcell from the period 2005 – 2014. Each figure shows a single UKESM1 gridcell, the monitoring sites (blue point) and the coastline (black solid line), where relevant.

We include this assessment in line 192 of the methods section:

"To further understand how successfully the site may represent the gridcell, we consider the homogeneity of the gridcells in terms of emissions and landcover. All urban sites will contain urban emissions sources, and some remote sites may contain biomass burning emissions. Using reanalysis data from GFED4s (Van der Werf et al., 2017), we confirm fire activity within the Daintree gridcell, in addition to some smaller and more infrequent burned areas at the Porto Velho and Yangambi sites. Furthermore, landcover in UKESM1 shows the Yangambi, Bukit Koto and Watukosek sites contain

some agriculture, and Bukit Koto, Watukosek, Daintree and Barro Colorado are all adjacent to ocean. Based on this analysis, the Amazonas and Santarem sites are the most homogeneous gridcells and therefore may be best represented by the model."

And the discussion (line 473):

"Several sites in this study are coastal (the model gridcell is split between ocean and land), namely Bukit Koto, Watukosek, Daintree and Yangambi. Due to a low deposition velocity of ozone over water (Sarwar et al., 2016; Luhar et al., 2018) and limited oceanic emission sources, concentrations of ozone over the ocean in UKESM1 are ~20 nmol mol^{-1}, and minimal diurnal variation is present. The gridcell chemistry and deposition velocities along coasts will be an average of the land and ocean, implying that the gridcell ozone concentration may not be representative of the site and the DOR is likely to be lower.

The resolution of UKESM1 can also introduce biases because emissions that, in reality, often occur as small, concentrated plumes are spread homogeneously across the whole gridcell volume. Of the remote sites included here, Daintree showed local fire emissions within the model gridcell that would be affected by this, possibly resulting in inaccurate representation of NOx concentrations and ozone formation."

For context, it is important to comment on how well UKESM1 represents surface ozone at midlatitudes and other regions outside the Tropics, and to provide a more critical assessment of how it compares with other global models in representing surface ozone (mentioned very briefly without detail on line 77). Reference to existing studies is fine for this, but the information is important for context, and the paper would be of wider interest if the issues are common to other models as well as just UKESM1.

We have added further detail to show that a positive systematic bias is a problem for several models in the tropics, although the magnitude of the bias varies. UKESM1 has one of the highest biases, although seasonality is captured better than several models. Thank you also for a reminder to compare to the midlatitudes. We have briefly provided context on line 434:

"In Europe and North America, UKESM1 tends to produce an underestimation of surface ozone in December-February and a positive bias in July-August (Archibald et al., 2020) similar to other models (Young et al., 2018). Turnock et al. (2020) found that recent earth system models have improved the negative bias over the Northern Hemisphere, but a positive bias remains elsewhere. Whilst the negative bias is attributed to excessive NOx titration, the cause of the positive bias has not been conclusively determined. Causes of model bias are discussed in more detail in Sect. 4.2 and 4.3.

This bias is not unique to UKESM1; a positive bias is present in several Earth System Models that took part in CMIP6, although it is larger than most in UKESM1 (Fig. S10). Crucially, however, the magnitude of the mean bias does not relate to the model's ability to capture the seasonal cycle, highlighting that bias in the mean state does not necessarily reflect the model representation of trends and variability. In this paper, we focus on UKESM1, confirming that there is a bias in the mean state in the tropics, yet also demonstrating the model has success in reproducing seasonality and the DOR at several sites. In this way, UKESM1 can be a useful tool to understand surface ozone processes and response to changing forcings. With appropriate bias correction, UKESM1 has been used to assess health burdens in different scenarios (Turnock et al., 2023; Akriditis et al., 2024) and this study allows further understanding of the bias in the tropics. This can reduce uncertainty in the assessments in this area."

With site-specific comments on line 398:

"In four other Earth system models performing the same simulation, the seasonality at Watukosek is captured better (Fig. S10f), whereas all models overestimate the change in ozone during biomass burning months at Yangambi, with UKESM1 performing among the best on account of the smaller seasonal variation (Fig. S10d)."

And line 534:

"Since other models with the same prescribed emissions display a seasonal cycle at Watukosek that looks closer to observations, the poor performance of UKESM1 likely relates to transport of anthropogenic emissions and their chemistry within UKESM1 rather than the emissions themselves."

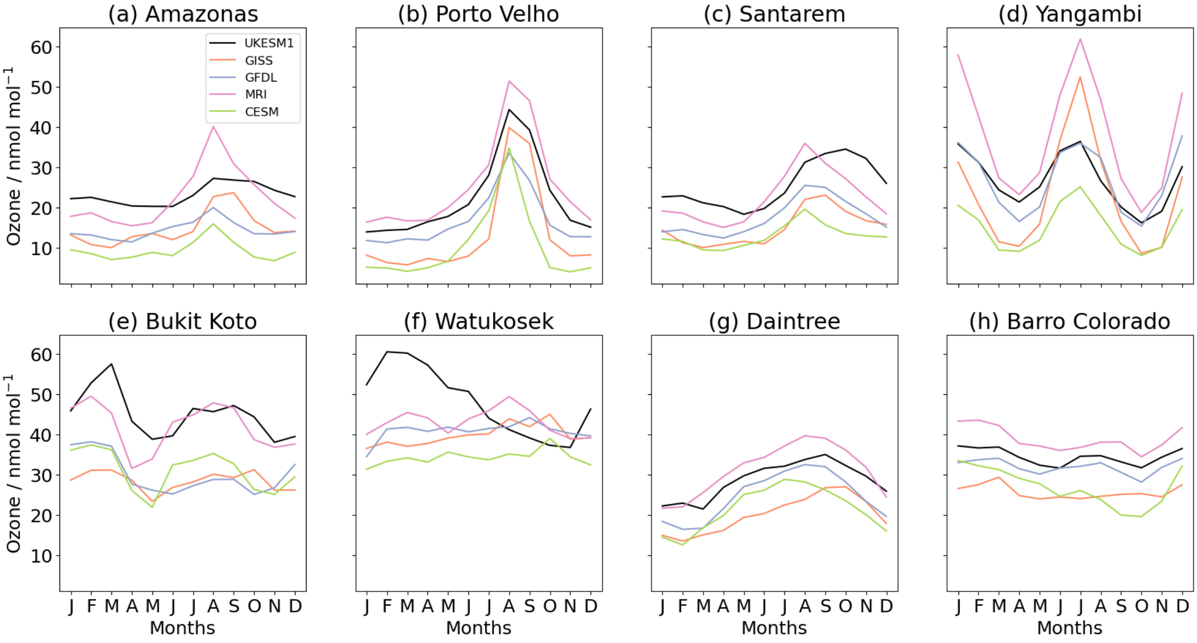


Figure S10: Monthly mean ozone for the period 2005 – 2015 at each remote site for five Earth System models for comparison to UKESM1 (solid black line).

Specific Comments

L.142: The 12 Tg/yr flux of soil NOx is as N, NO or NO2? If this flux is quantified, it would be useful to give the average or range of the lightning NOx source, and perhaps also the bVOC emissions.

Thank you for pointing this out and we agree these values are a helpful addition to the text (line 153).

"Lightning NOx is calculated using the parameterisation of Price and Rind (1992), which calculates a lightning flash density based on cloud-top height and produces a global annual emission rate of 5.93 Tg-N yr^{-1} over 2005 to 2014. Soil NO is prescribed as a spatially explicit model output according to Yienger and Levy (1995), scaled to give an annual flux of 12 Tg-NO. CH_4 is prescribed as annual mean surface concentrations based on observations over the historical period (Meinshausen et al., 2017). Emissions of isoprene and monoterpenes are generated by the interactive biogenic VOC (iBVOC)

emission model (Pacifico et al., 2011) with annual mean emissions of 495.9 Tg-C yr^{-1} and 115.1 Tg-C yr^{-1}, respectively."

L.147: This sentence belongs in the first paragraph of the section, before the discussion of emissions. It would also be helpful to indicate the level of complexity represented in the NMVOC chemistry (just longer-lived VOC and isoprene, or some treatment of other reactive VOC?)

We have rearranged this sentence and added more specific details.

L.163, 165: it is distracting to the reader to see references to Figures later in the results section while still in the methods; please remove these and present the analysis information in a more generic way here. Detail relevant to a specific figure should be included where that figure is presented.

We have removed reference to individual figures.

L.205: clarification needed: does the standard deviation here quantify the interannual variability (or the seasonal variability)?

Text has been edited to qualify this.

L.299: A p-value of this magnitude suggests incorrect application of statistical methods.

Thank you for noticing, we have revised the calculation to produce the correct number (0.002, line 342).

L.317: Subtitle better as "How well does UKESM1 reproduce...."

Thanks for your suggestion, which we have taken on board.

L.320: It is not clear what "pattern" refers to here (diurnal, seasonal, or spatial variation?)

We have corrected the unclear wording used previously.

L.337: model resolution is not a cause of model bias, it is a structural characteristic which leads to biases from the processes that are represented.

Thank you for this clarification, we have edited the text to be more precise with our language.

L.381: interactive natural precursor emissions probably are very important for reproducing tropical ozone, but your studies haven't shown this. A simple sensitivity study would allow you to confirm (and quantify) this effect.

We accept that this is not confirmed or tested within the study, so we have removed the comment. This comment served as a possible explanation for why daily variability in O3 concentrations are observed despite monthly mean precursor emissions. Since the model meteorology it not nudged to observations, and often the data is limited, evaluation of the model at the daily level is difficult and we can only look at the distribution. Sensitivity of ozone to BVOCs and their chemistry in the tropics has been considered previously by Brown et al. (2022) and Weber et al. (2023).

L.453: resolution is not an attributable cause of bias, please rephrase here.

We have rephrased.

L.457: It is clear that surface ozone from global models such as UKESM1 should not be used for health and ecosystem impact assessments at the current time. Substantial revision of this point is required.

If the mean bias is taken into account, we believe UKESM1 can be a valuable tool. Indeed, bias corrected health impact analysis have already been published. The statement has been amended to address the model limitations. This is discussed in a previous response to reviewer 1:

"Analysis of the DOR allows local-scale responses to be considered separately to the systematic bias and may be a useful diagnostic for other researchers to consider. Overall, our results suggest that UKESM1 may be useful for understanding ozone responses to forcings but hourly data should not be used 'off the shelf' for health and ecosystem impact assessments. Bias correction is an option to avoid overestimation of the risks but users should be aware that monthly mean concentrations may require multiplicative bias correction in biomass burning regions and that gridcells containing non-homogeneous emission sources or land cover types may be impacted by the negative effects of coarse model resolution more than pristine regions. The magnitude of the bias in different regions and seasons, and its dependence on factors such as distance from emissions sources remains to be quantified. For this, more in situ monitoring is instrumental. "

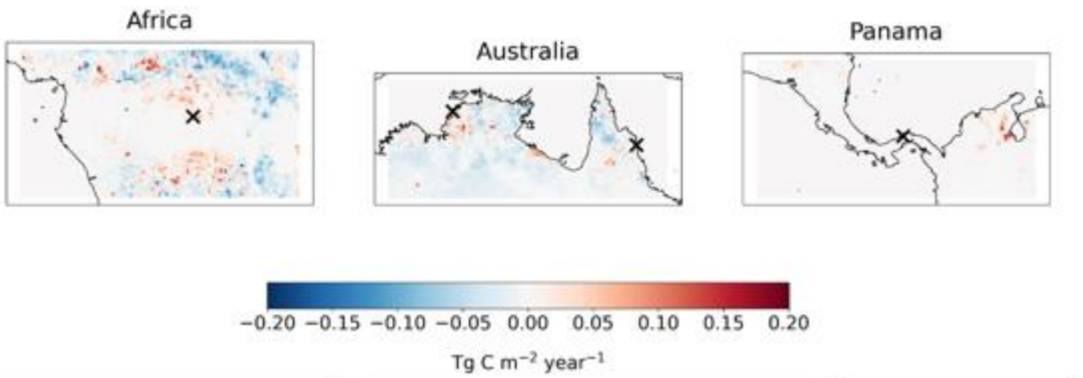
Fig S1 is cramped, although would be more readable if included in a vector format rather than as an image. The blue lines would be more accurately described as green (although the overlay on yellow bars makes this difficult to discern).

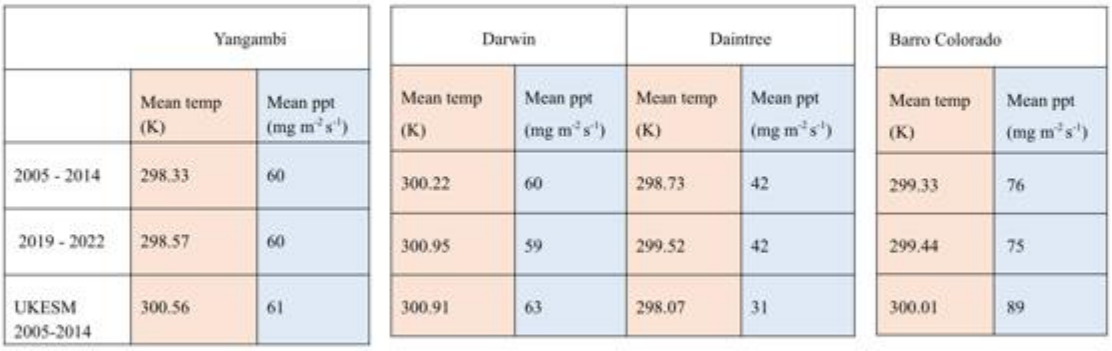
We have spread this figure over two pages to improve readability. The final figures will be submitted in an appropriate format for the journal.

Fig S3: In light of the points about temperature sensitivity raised in section 2.3, it would be useful to include the mean temperatures and precipitation from UKESM1 in the table component to highlight how this compares with the reanalysis.

This is a great suggestion and we have added a comparison starting on line 174:

"Comparison to reanalysis shows UKESM1 overestimates annual mean surface temperature by an average of 0.7 K in the period 2005 – 2014 (Table S2). Archibald et al. (2020) show that the temperature sensitivity of ozone in the chemistry scheme of UKESM1 is on the order of 1 nmol mol^{-1} K^{-1} in the absence of feedbacks from the land surface, meaning climate trends are unlikely to cause a significant difference in ozone between the different periods. Differences in observed climate at sites where the model and observation period are mismatched are given in Fig. S3 and show the model period (2005 – 2014) differs from the observation period (2019 – 2022) by 0.5 K on average using reanalysis, and that UKESM1 temperatures are closer to those observed in 2019 - 2022."





	Yangambi		Darwin		Daintree		Barro Colorado	
	Mean temp (K)	Mean ppt (mg m^{-2} s^{-1})	Mean temp (K)	Mean ppt (mg m^{-2} s^{-1})	Mean temp (K)	Mean ppt (mg m^{-2} s^{-1})	Mean temp (K)	Mean ppt (mg m^{-2} s^{-1})
2005 - 2014	298.33	60	300.22	60	298.73	42	299.33	76
2019 - 2022	298.57	60	300.95	59	299.52	42	299.44	75
UKESM 2005-2014	300.56	61	300.91	63	298.07	31	300.01	89

Figure S3: The change in fire emissions between 2005 – 2014 and 2019 – 2022 in the area around the Congo site, the Darwin and Daintree sites and the Panama site. Tables below show the mean temperatures and precipitation rates for the grid cells containing each site using CRU-JRA reanalysis at the same resolution as UKESM1 and UKESM1 for the period 2005 – 2014.

Fig S8: If site T1 does not have data between April and December then these points should not be joined on the graph.

We have amended this.

Technical Corrections

There are some weaknesses in writing style in places that could be removed with a thorough proof-read. Other minor points:

Table 1: lat/lon of grid cell centers are not needed to four decimal places (one would be sufficient)

This has been changed.

L.115: Table 1 indicates that Sao Paulo spans three gridcells, but the text states two; please correct text or table as appropriate.

Thank you for identifying this typo. It has been corrected.

L.156: ppb used here, but nmol/mol used elsewhere

This has been changed.

L.192: add units: nmol/mol after 13.0. Note also that the percentage biases in the following sentence can't justify quotation to one decimal place.

This has been changed.

L.213: trends conventionally refer to changes; rephrase here (also at L.266)

Thank you for highlighting this.

L.279: "factors rather than..." meaning of this sentence unclear, please rephrase

The paragraph (line 322) has been rephrased to read:

“The bias at the Bukit Koto, Yangambi and Porto Velho sites may be amended using a multiplicative linear correction rather than an additive correction because the modelled seasonal cycle has greater monthly variability in ozone than the observations. Applying a bias correction multiplier of 0.33, 0.55 and 0.25 for Bukit Koto, Yangambi and Porto Velho, respectively, brings the magnitude of the monthly means and the seasonal variation closer to observations (Fig. S6), however it is not necessarily suitable for correcting daily or hourly biases. The seasonal variability at these sites is dominated by changes in biomass burning, suggesting that the model overestimates ozone formed from burning due to either incorrect emissions or process representation. At the other remote sites, the bias is consistent between months and therefore the annual means in Fig. 2 represent the biases sufficiently well, or the seasonal cycle is not well represented. In these cases, scaling the model output as in Fig. S6 removes the seasonality, suggesting a background bias that is not dependent on the local ozone concentration.“

L.441: "across" not needed

This has been corrected.

L.459: "and is"

This has been corrected.

Figs 2, 3: x-axis interval in right panel should be 10 for consistency with y-axis.

This has been corrected.

Reviewer 2:

In this manuscript, the authors evaluate surface ozone simulated by UKESM1 in the Tropics. The topic of the manuscript is scientifically very relevant for for the community but I wish the authors enhance their process understanding. They use a number of statistical methods to quantify the model performance. However, often the interpretation comes too short and thus all in all many aspects like the systematic bias of UKESM1 remain unexplained.

To compare station measurements with zhe model (grid cell) output the authors average the measurement data within one grid box. I think this is inaccurate and introduces uncertainty, since the stations have, as they mention later, different meteorological conditions. At urban stations, in particular, a complex chemistry at sub-grid scale occurs and thus they are hard to represent at coarse spatial resolution of ~140 km like here.

We completely agree with the reviewer on his argument that cross-gridbox averaging is an important source of uncertainty in model-to-obs comparison. However, at coarse model resolution averaging over stations within a gridbox with varying chemical and meteorological conditions is in fact closer to what the model does. The model represents the “mean state”, the average over a large area with one value. The assumption is that by applying a similar approach to the obs we bring model and obs closer together in their statistical meaning. That is also why dedicated model sensitivity studies are required to assess the impact of model resolution on model performance, which go beyond this study.

The method we use is in fact standard procedure and is the typical way in which sites are evaluated in comparison to models (see previous studies by Gaudel et al. (2018); Young et al. (2018) Griffiths et al. (2021) as part of TOAR I). Since every model should be similarly affected by this issue, it makes multi-model comparison possible. As we cannot reduce the gridbox size of this simulation, we must hope that by averaging all the measurements within the gridcell we get closer to the representation of the whole gridcell. The more heterogeneous the gridcell, the less true this is. For this reason, the analysis focuses mainly on remote sites, which sample a bigger footprint than urban sites with much steeper spatial gradients, and the challenges and limitations of model resolution are described in detail. See an earlier comment about which gridcells may be considered homogeneous vs heterogeneous in our study.

An interesting conclusion is that urban areas do not significantly show different biases at an annual scale, despite these known issues with model representation. Since there are very limited measurements of tropical O3, with most put in place to monitor urban pollution, we felt these sites were useful to include to provide greater context and as much data as possible. However, the text quickly moves to focus on remote sites.

Gaudel, Audrey, et al. "Tropospheric Ozone Assessment Report: Present-day distribution and trends of tropospheric ozone relevant to climate and global atmospheric chemistry model evaluation." *Elem Sci Anth* 6 (2018): 39.

Young, Paul J., et al. "Tropospheric Ozone Assessment Report: Assessment of global-scale model performance for global and regional ozone distributions, variability, and trends." *Elem Sci Anth* 6 (2018): 10.

The authors apply UKESM1 in a free-running mode, as far as I understand. From my knowledge, this is very uncommon for a model evaluation/ comparison with measurements. Does the model represent the meteorology in the studied regions realistically?

We acknowledge that in some cases a nudged model is used, for example to test a newly developed chemsitry scheme. In other cases, all the features and feedbacks are evaluated using a free0running model. Here, we have chosen the free-running simulation as this was the simulation for which hourly data was available. To record hourly output is storage intensive and adds a great amount of simulation time in order to write the data, so it is not standard or widely available and CMIP6 does not provide the nudged simulations that are requested.

Our choice to use CMIP6 data also allows for multi-model comparison and comparisons to other literature that use CMIP data (Young et al. (2018); Griffiths et al. (2020)). There are many more studies that make use of the same dataset which use model-to-obs comparison to evaluate model performance. Additionally, data users who may be interested in using freely available data, e.g. for impact assessment, are likely to use the output of the free-running model (e.g., Turnock et al., 2023; Akridis et al., 2024). Our analysis serves as validation of and guidance for bias correction of these studies.

A climate model represents the mean climatological state and by using climatological means of the obs we compare like for like as best as is possible when comparing models with site level observations. A comparison table of temperature is now included as we agree this is helpful information for the reader.

Line 174: *"Comparison to reanalysis shows UKESM1 overestimates annual mean surface temperature by an average of 0.7 K in the period 2005 – 2014 (Table S2)."*

Table S2: The gridcell mean and standard deviation in temperature for the period 2005 – 2014 from UKESM1 and CRU-JRA reanalysis. The final column gives the difference between model and observations.

Site	CRUJRA mean	UKESM1 mean	CRUJRA std	UKESM1 std	Difference (UKESM1 – CRUJRA)
	°C	°C	°C	°C	°C
Porto Velho	20.9	22.0	0.2	0.6	1.1
Amazonas	22.1	23.4	0.2	0.6	1.3
Santarem	21.9	23.1	0.1	0.7	1.3
Yangambi	19.3	21.6	0.1	0.4	2.2
Bukit Koto	20.9	21.1	0.2	0.2	0.2
Watukosek	20.5	21.6	0.2	0.3	1.1
Daintree	19.7	19.0	0.4	0.3	-0.7
Barro Colorado	20.3	21.1	0.2	0.3	0.8
Bogota	13.9	14.7	0.3	0.4	0.8
San Lorenzo	17.4	16.8	0.4	0.8	-0.6
Sao Paulo	13.9	14.3	0.3	0.3	0.4
Jakarta	20.4	21.6	0.1	0.2	1.2
Darwin	21.2	21.9	0.6	0.4	0.7

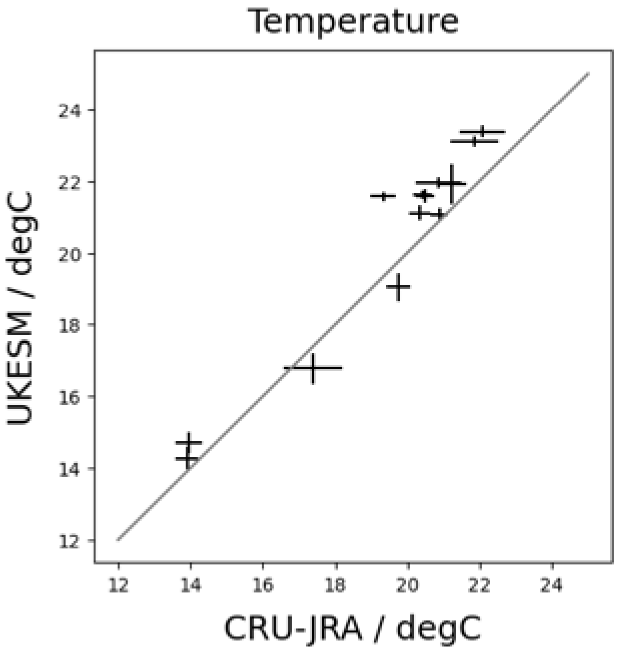


Fig (not in manuscript): Relationship between annual mean temperature from reanalysis (CRUJRA) and UKESM1 at each site and the 1:1 line (grey solid line). Bars show one standard deviation in the annual means.

Young, Paul J., et al. "Tropospheric Ozone Assessment Report: Assessment of global-scale model performance for global and regional ozone distributions, variability, and trends." *Elem Sci Anth* 6 (2018): 10.

Griffiths, Paul T., et al. "Tropospheric ozone in CMIP6 simulations." *Atmospheric Chemistry and Physics* 21.5 (2021): 4187-4218.

Also, the usage of the diurnal ozone range for showing the model's feasability to capture the diurnal varaition of surface ozone, does not convince me.

It shows that the model is able to capture the magnitude of the increase over the day. It is true that there are other features to the diurnal cycle, such as the shape. We show the shape of the diurnal cycle compared to observations in Fig. S4. We also show the timing of the minimum and maximum in Fig. 4. Using all these data we feel the diurnal cycle is captured well. We then choose to use the diurnal ozone range as a quantitative metric to allow for more statistical analysis, that we feel is informative.

We discuss in detail the difference between the DOR and monthly mean O3, thereby highlighting the benefits of the DOR, in response to the first general comment by reviewer 1.

Minor comments:

The description of VOC/NOx-limited regimes could be more clear (l. 63 ff).

We have added further information to line 63:

"*The ozone production rate is controlled by the reaction NO + HO2/RO2 and can therefore be considered NOx-limited or VOC-limited depending on the availability of these species (Archibald et al., 2020b; Wild and Palmer, 2008). However, effect of changing NOx and VOC concentrations on ozone concentrations is non-linear. For example, in a VOC-limited regime, reducing NOx concentrations will not decrease the rate of ozone production (and likely increase the rate).”*

l. 72: 'good test space' -> 'much study potentilal'

We have changed this phrase.

l. 145: Why don't you use the more recent estimates by Sinderalova et al. 2022?

These simulations were undertaken for CMIP6, before the data by Sindelarova was published. However, since isoprene and monoterpenes are produced interactively, the prescribed BVOCs likely account for a small fraction of O3 chemistry so updating is unlikely to make a significant difference. We use CMIP6 data, which enables comparison to other literature (both of model evaluation and impact assessments), and we feel this is more beneficial than updating to the newest dataset.

What is the global annual lNOX emission?

This has been included in the methods section.

Don't you think that enhanced process understanding and using a more complex SOA chemistry (for example) could resolve some of these model biases? Can you state whether the biases are model-specific and why? (l 385 ff.)

Yes enhanced process understanding surely could resolve some of these model biases, SOA chemistry being one option. Some other processes include missing ocean chemistry (e.g., halogen chemistry) or deposition to reduce O3 over the ocean, missing in-canopy chemistry, for example sesquiterpene chemistry or just the fact that the canopy stability can separate the trunk space from the air above, leading to high O3 destruction by BVOCs in a dark environment with low photolysis. Fire plumes could be better represented chemically as the highly concentrated plume followed by aging is not represented or well understood. As mentioned within the text, more explicit treatment of isoprene chemistry has been tested by Weber et al. (2020) and the bias increased as a result, but this only tells us that there is even more work to be done, likely a combination of all of these processes that combine non-linearly. In short, we do not believe there is one quick fix that can be tested within this paper.

On the subject of whether the bias is model specific, we have now included a comparison against other models (see previous response) and in previous work (Brown et al., 2022). It is interesting that the magnitude of the mean bias is unrelated to the model ability to reproduce the seasonal cycle. This indicates that the background O3 (the mean state) should be considered separately from variability, which is the approach we take in this study. Evaluation in our previous work finds many differences in chemistry and representation between models; one feature that stands out is variation in surface NOx concentration between models. For example, the GISS model has much lower surface NOx concentrations, perhaps because of the much larger vertical levels that allow NOx to escape more easily, in addition to much higher isoprene emissions that remove NOx. However, this does not come close to resolving why some models have a larger bias than others. For a detailed inter-model comparison, we direct the reader to literature aimed at investigating this question (Young et al., 2018; Griffiths et al., 2021).

Weber, James, et al. "Improvements to the representation of BVOC chemistry-climate interactions in UKCA (vn11. 5) with the CRI-Strat 2 mechanism: Incorporation and evaluation." *Geoscientific Model Development Discussions* 2021 (2021): 1-52.

Brown, Flossie, et al. "The ozone–climate penalty over South America and Africa by 2100." *Atmospheric Chemistry and Physics* 22.18 (2022): 12331-12352.

Young, Paul J., et al. "Tropospheric Ozone Assessment Report: Assessment of global-scale model performance for global and regional ozone distributions, variability, and trends." *Elem Sci Anth* 6 (2018): 10.

Griffiths, Paul T., et al. "Tropospheric ozone in CMIP6 simulations." *Atmospheric Chemistry and Physics* 21.5 (2021): 4187-4218.