Dear editor:

We thank you for your comments to further improve our manuscript, and address your concerns below. For clarity, the comments are given in bold, followed by our responses. The modified text in our revised manuscript is given in italics and blue.

I will accept the manuscript subject to minor revisions. Reviewer 2 suggested a rejection of this manuscript, as using the CAMS model as a reference case was subject to substantial biases (essentially the model is not strongly constrained by observations at the surface). I do not concur with the reviewer 2's recommendation, and motivate this as follows.

In your revised manuscript you convincingly demonstrate there is nevertheless a substantial merit in using CAMS as the observational 'truth'. This is because the biases in the UKESM1 are substantially higher than in CAMS.

We thank you for your recognition of the merit of our work and the revised version of the manuscript.

However, it would be good to highlight in abstract, and conclusions that there is also a risk of over-interpretation when using another model for the machine learning based bias correction, when essentially similar processes may have lead to biases in the models, output fields are more smooth than in the real-world etc.

We thank you for pointing out that using non-observation based data as reference may still lead to biases in explaining deep learning outputs. We did this because of its relatively small biases in surface O_3, as you mentioned above, which indeed provides a great opportunity to correct biases for other models with large errors such as UKESM1. However, we acknowledge that the systematic errors due to the issues in processes and spatial representativeness may occur in all physical models, and cannot be fully eliminated, which may impact the bias correction. We have now acknowledged this point in both the abstract and conclusions.

Page 1, line 9:
... Atmospheric chemical species such as the hydroxyl radical, nitric acid and peroxyacyl nitrate show strong positive relationships with ozone biases on a regional scale. These relationships reveal the conditions under which ozone biases occur, although they reflect association rather than direct causation.

Page 17, line 361:
... We also note that there are weaknesses in the representation of O_3 in the reanalysis data which are likely to affect the magnitude of the biases we have derived. However, we have successfully demonstrated the feasibility of bias correction using this data, and will explore the challenges of data sparsity and spatial representativeness associated with use of surface measurements directly in future work.

The conclusions at line 333-336 already gives some hints on how you want to address these issues, but need to be expanded to better indicate what the limitations of the study are. Also please discuss to what extend the method is applicable for other models which are less biased (i.e. similar magnitude of biases as CAMS).

We thank you for the suggestions. Explaining machine learning outputs is always difficult because it is based on statistics, and that is where the limitation of the study is. It is difficult to use a few variables to identify the specific processes that have the largest biases, and it is also a challenge. We have acknowledged the limitations in the Conclusions section, and discussed the method that may be more suitable for models with small biases.

Page 17, line 359:

... However, we demonstrate that the relationships between the variables with the highest feature importance and surface O_3 biases are intuitive, e.g. with temperature and photolysis rates, and this provides useful insight for further model improvement. While we are not able to identify the specific processes leading to biases using this approach, it allows us to target processes that are most sensitive to these variables. It would be valuable to develop explainable machine learning algorithms to use for bias correction.

Page 17, line 364:

... However, we have successfully demonstrated the feasibility of bias correction using this data, and will explore the challenges of data sparsity and spatial representativeness associated with use of surface measurements directly in future work. This approach should also be directly applicable for models with smaller initial biases, and in this case it would be particularly valuable to consider daily or hourly mean O_3 to explore representation of synoptic and diurnal variations in O_3. However, the development of a robust and reliable surface O_3 climatology based on observations would be particularly useful to improve assessment of model biases.