Author's comments in reply to the anonymous referee for "Identifying Drivers of Surface Ozone Bias in Global Chemical Reanalysis with Explainable Machine Learning" by Miyazaki et al

Reply to Referee #2

The authors use an RF algorithm to (1) emulate predicted concentrations of surface ozone from a leading tropospheric chemical reanalysis product, and (2) predict the bias of the reanalysis product relative to global surface observations. The authors use explainable AI techniques to understand drivers of bias in ozone reanalysis. These results offer a useful perspective on O_3 chemical transport model predictions and data assimilation output, and I recommend publication after the following comments are addressed.

Major points:

- *My biggest concern by far in this work is spatial extrapolation: the TOAR surface data used in training are clustered in a few regions (North America, Europe, and east Asia) while bias is predicted globally. The authors are aware of this, but spatial crossvalidation is a more direct way of quantifying the issue and is not done in this work. I am most concerned about (1) oceans, (2) boreal regions, and (3) the tropics where training data is limited. Consider withholding some of the few training sites in these regions and measure how well the RF predicts bias there (the clustering maps in Figure 12 might be a reasonable way to do crossvalidation). The authors could also use more recent observations in China and India for independent evaluation. Do we really have enough data to use RF, a highly data-dependent algorithm, for extrapolation to these regions?*

We thank the reviewer for raising this important concern regarding spatial extrapolation. We fully agree that evaluating the model's ability to generalize beyond observationally dense regions is critical for assessing the robustness of our approach. In response, we have taken the following steps and provided clarifications:

1. **Assessment through emulator runs:**

To examine the impact of sparse observational coverage, we performed emulator experiments in which the ML model was trained only on TOAR-covered regions and evaluated globally. These results revealed that:

- Prediction errors were systematically higher in the tropics, oceans, and high-latitude boreal regions compared to the NH midlatitude regions with dense TOAR coverage.
- Oceanic regions exhibited particularly large uncertainties, consistent with the absence of surface constraint.

- These outcomes highlight the degradation of predictive skill in observationally sparse regions and demonstrate the intrinsic limitations of spatial extrapolation in such settings, even in the absence of formal spatial cross-validation.

2. **Justification of temporal cross-validation:**

In the ML runs using actual observations, we employed a leave-one-year-out **temporal cross-validation** strategy rather than spatial cross-validation. This decision was based on the sparsity and irregular distribution of TOAR sites, which made it difficult to define spatially independent and statistically robust validation subsets. Spatial cross-validation under such conditions would risk creating training or validation sets that are too small or not representative of broader spatial patterns. While this limits our ability to assess true spatial extrapolation, temporal cross-validation allows us to evaluate the model's generalizability across time, which is still relevant for long-term bias analysis.

We now explicitly discuss this in the revised manuscript in Section 2.3, where the cross-validation strategy is described as follows:

"While the temporal cross-validation approach, implemented through a leave-one-year-out strategy, does not fully address the challenge of spatial extrapolation, it provides a robust framework for evaluating the model's generalization across years with diverse chemical and meteorological conditions. We acknowledge that spatial cross-validation would offer a more direct assessment of the model's extrapolation capability. However, this was not feasible in our case due to the sparse and uneven distribution of TOAR monitoring sites, particularly outside of North America, Europe, and East Asia, which results in limited spatial coverage and strong regional clustering. In many under-sampled regions, such as the tropics, boreal zones, and the Southern Hemisphere, the lack of contiguous observational clusters prevents the construction of spatially independent and statistically meaningful training and validation sets. Consequently, we relied on temporal cross-validation to preserve both data representativeness and model stability, while recognizing that spatial extrapolation remains an important area for future investigation. To complement this limitation, the ML's predictive performance in observationally sparse regions is further evaluated through dedicated emulator experiments described in the following section."

3. **Limitations of Random Forest for extrapolation:**

We also acknowledge the fundamental concern raised by the reviewer regarding the suitability of RF for extrapolation. As an ensemble method, RF tends to interpolate within the convex hull of the training data and is not inherently designed to predict beyond regions with observational coverage. To clarify this limitation, we have added the following statement to Section 5.1:

"In addition to the data imbalance, RF itself has inherent limitations in extrapolation. As an ensemble tree-based method, RF primarily interpolates within the convex hull of the training data and lacks the ability to generalize to regions with little or no observational coverage. Consequently, predictions over sparsely observed areas, such as the tropics and oceans, are subject to greater uncertainty and should be interpreted with caution. These algorithmic and data-related constraints underscore the need to expand global monitoring networks and explore hybrid approaches that integrate physical knowledge with ML."

4. **On the use of recent observations in China and India:**

We appreciate the reviewer's suggestion to incorporate more recent observations from China and India for independent evaluation. At the time of this study, these data had not yet been fully integrated into the TOAR database. We agree that including such data would enhance the spatial representativeness of the training set and reduce extrapolation bias. We have addressed this point in the revised manuscript with the following statement in Section 5.1:

"In addition, surface ozone observations from emerging monitoring networks, including those in China and India, were not yet fully incorporated into the TOAR database at the time of this study. Their inclusion in future work is expected to improve spatial representativeness, reduce extrapolation bias, and strengthen the reliability of ML-based inference in currently under-sampled regions."

- *In cases of highly imbalanced training sets, where some regions are far overrepresented, methods like SMOTE or weighted training are sometimes employed to ensure that the RF is penalized more heavily for bad predictions at some sites. Did the authors consider using such approaches?*

Thank you for this valuable comment. We acknowledge that the spatial distribution of TOAR sites is highly imbalanced, which may introduce regional biases and limit the model's ability to generalize to globally diverse conditions. While SMOTE is designed primarily for classification problems, techniques such as weighted training or stratified sampling can be effective in addressing imbalance in a regression setting. In the present study, we prioritized model interpretability and transparency, and therefore did not implement these strategies. However, we agree that this is an important direction for future work,

particularly as more globally representative observational data become available. To reflect this point, we have added the following sentences to the discussion section (Section 5.1), highlighting the potential role of weighting methods:

"We also note that the spatial distribution of training data is highly imbalanced. This imbalance may lead to an overrepresentation of region-specific patterns in the learned relationships, potentially limiting model generalizability. Although we did not implement weighting or rebalancing strategies such as region-based sampling weights or stratified training in this study, such techniques may offer an effective means of mitigating spatial biases in future applications."

- *Explainable AI methods are vulnerable to collinearity in the inputs, as the authors are aware, and of the algorithms used SHAP (TreeExplainer) is most robust to this problem. I would like to see more comparison between SHAP and the other methods. For example, in Figure 6, what does SHAP suggest are the top contributors to ozone bias in these regions? In the literature, for SHAP regional attribution some authors use separate RFs trained on training sets focused on particular regions.*

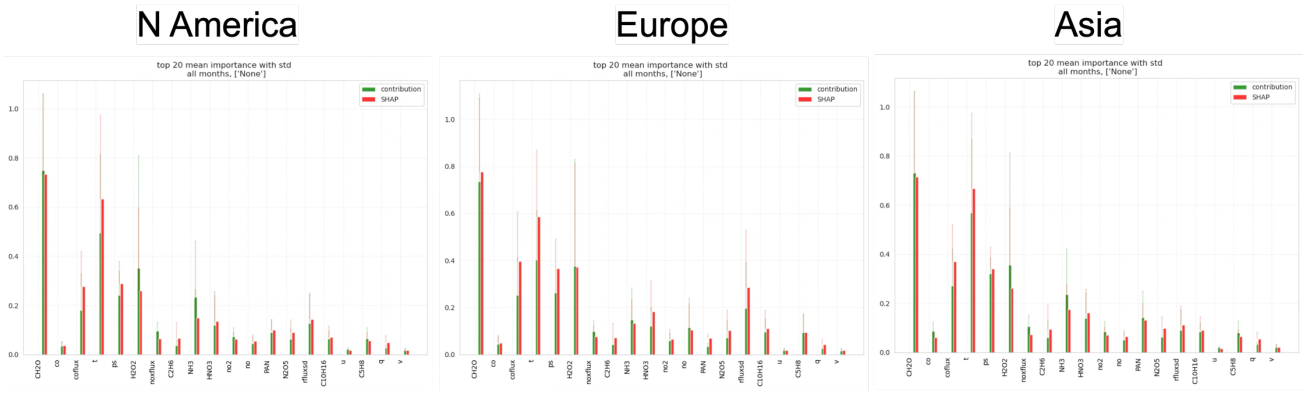

Fig. Comparison of feature attribution metrics derived from SHAP and Conditional Feature Contributions (CFC). Bars represent normalized contributions of the variables to the surface ozone bias.

Thank you for this insightful comment. As shown in the comparison figure, the feature attribution metrics derived from SHAP and Conditional Feature Contributions (CFC) are highly consistent across regions. This agreement is expected for tree-based models such as RF, where both methods yield additive explanations by decomposing predictions into contributions from individual input variables. While SHAP uses a game-theoretic framework that averages over all possible feature combinations, CFC calculates importance based on the actual traversal paths of each instance through the decision trees. In settings with moderate collinearity and limited feature interactions, the two approaches tend to produce similar results

(Lundberg et al., 2020). The strong alignment observed in our study supports the robustness of the derived feature importance rankings. We have added a brief discussion of this comparison in Section 4.1 of the revised manuscript.

"We also compared the feature attribution results with SHAP values (not shown). In particular, SHAP and CFC yielded closely aligned rankings of the dominant contributors to surface ozone bias across regions. This agreement is expected for tree-based models, as both SHAP's game-theoretic averaging and CFC's path-based decomposition provide additive explanations. The consistency between these two approaches, especially under conditions of moderate input collinearity (Lundberg et al., 2020), supports the robustness of our feature importance analysis."

Minor points:

- *Figure 3: It is not surprising that RF has trouble predicting the distributional tails; this has long been observed in the literature (and is to be expected given it is an ensemble algorithm). Consider commenting on the limitation of this method for e.g. improving models such that they give better predictions of NAAQS ozone exceedances (e.g. MDA8).*

We agree that RF models are known to underperform in capturing the extremes of the target distribution, largely due to the averaging nature of ensemble predictions, which is a limitation well documented in the literature. This tendency is also evident in our results. To clarify this point, we have added a short discussion in Section 5.2 along with citations to relevant studies (Gao et al., 2022; Chen et al., 2021) that report similar findings.

"The ML predictions systematically underestimate the variability in surface ozone bias across all regions, indicating an underestimation of the occurrence of extreme (both positive and negative) bias values. This behavior is a well-known limitation of RF, which tend to underpredict distributional tails due to their ensemble averaging structure (Gao et al., 2022; Chen et al., 2021). Such underestimation is particularly relevant when aiming to detect exceedances of air quality standards, where accurate representation of high-ozone events is critical."

- *Given the given tropical Pacific pattern in RMSE (Figure 2) I am curious about the role of ENSO in driving RF error. Is lightning NOx a problem here?*

We have not explicitly evaluated interannual variability or the influence of ENSO in this study, so its role in driving RF prediction errors remains unclear. In addition to lightning NO_x, ENSO-related changes in

convection, cloud cover, and large-scale atmospheric circulation could indirectly affect surface ozone and contribute to the observed regional error patterns. Investigating these mechanisms, particularly through targeted analysis of interannual variability, represents a valuable direction for future research.

- *Could you clarify if TOAR surface sites averaged to the grid of the TCR-2 output? In places with many monitors within a single grid cell this could lead to sample bias where e.g. urban areas are even more disproportionately represented.*

Yes, TOAR surface ozone observations were aggregated to the 1.125° × 1.125° grid used in the TCR-2 reanalysis. In grid cells containing multiple monitoring sites, we used the median of all available TOAR observations to represent the surface ozone value, thereby reducing sensitivity to outliers and localized effects, particularly in urban environments. In regions with dense urban monitoring networks, this aggregation may introduce sample bias by over-representing urban conditions relative to the true grid-scale chemical environment. While this limitation cannot be entirely eliminated given the current observational distribution, we now address this issue explicitly in the revised manuscript. The following statement has been added to Section 3.2:

"In particular, spatial smoothing resulting from the relatively coarse resolution of the reanalysis can limit the ML model's ability to capture fine-scale chemical and dynamical processes, especially in urban environments. The aggregation of urban and non-urban chemical regimes within individual grid cells can introduce representativeness errors that add uncertainty to ML predictions. Depending on the magnitude and spatial variability of sub-grid processes, this may lead to systematic underestimation or overestimation of the reanalysis bias."

- *Figure 5: consider using same colorbar for observations and for predictions.*

Applied

- *Figure 11: consider also showing uncertainty as percentage of predicted bias*

Thank you for the suggestion. We considered showing uncertainty as a percentage of the predicted bias. However, in regions where the predicted bias is small, the relative uncertainty becomes disproportionately large and may misrepresent the true model behavior. For this reason, we chose to present the uncertainty in absolute terms, which more consistently reflects the magnitude of prediction spread across all regions.

- *Throughout, increase font size of figures. It can be quite hard to read.*

We have increased the font size in some figures to improve readability.

- *Some typos throughout. Here are a couple: Line 83: "the simulation of simulate" should read "the simulation of". Line 241: Missing unit after "exceeded 30" (I think it should be percent)*

We have carefully re-checked the manuscript and corrected some typos.