We thank both referees for the obvious time and care they put into their reviews, which helped us to revise the manuscript with improved focus and clarity. We have addressed all of the referee comments as described below. In addition, the figures and results were completely revised due to an error that we recently discovered in the particular ozone product that we retrieved from the TOAR Surface Ozone Database and used in this analysis. The product was the monthly mean of the maximum daily 8-hour average (DMA8), calculated for each site in the TOAR database. Close inspection of the product and comparison to daily DMA8 ozone values at individual sites revealed that the sampling of the daily DMA8 values for this particular product was in error, which resulted in monthly means that were biased high. As a result the observed 6-month running mean of the monthly mean DMA8 values used in this analysis was biased high by approximately 25%. The error has been corrected in the TOAR database and in the archived TOAR data products. This analysis has also been updated with the corrected data. However the method for constructing our final fused surface ozone product (M^3Fusion) did not change. The final corrected product shows that the atmospheric chemistry models are generally biased high with regards to the 6-month running mean of DMA8. As described in the new concluding paragraph at the end of the manuscript, this is an important result which demonstrates the usefulness of our method for bias correcting model output.

Anonymous Referee #1

We thank the reviewer for providing valuable comments on our manuscript. The reviewer comments are shown below in bold font, followed by our response in normal font.

This manuscript presents a new statistical method for combining observations of surface ozone with model outputs. The manuscript is clearly written and the method is well described. The fused data set represents a significant output that could be useful to analyze the relevance of ozone to health impacts.

The manuscript is nearly ready for publication, but I have several questions and editorial suggestions for the authors, listed below.

1. I suggest to combine Section 2.2 and Section 2.3 into one. Section 2.3 describes the implementation details but ends up repeating concepts already described in Section 2.2, resulting in poor readability.

Thanks for the suggestion. We have merged these two sections and removed the overlapping concepts.

2. To create the interpolated field from ozone observations the authors used a Bayesian approach that allows for the quantification of the uncertainty in the gap-filled product.

2.1. Can the authors comment on why they choose not to account for the sampling uncertainty, even though it could be easily estimated from the posterior?

Accounting for the sampling uncertainty in the data fusion process is a difficult task and according to the referee's comments we now include some discussion of this topic at the end of Section 2:

“We adopted a regression weighting approach that only accounts for the mean spatial fields of the interpolated ozone and model output, rather than the underlying associated uncertainty. We take this approach due to the prohibitive size of high resolution output (over 1 million output points for each model), but also due to the lack of a thorough investigation regarding the ideal method for combining models based on different sources of uncertainty. For example, the interpolation uncertainty can be quantified easily through the posterior distribution and considered to be related to measurement error (small scale) or sparse sampling across a region (large scale), however, model uncertainty is a different concept altogether that could result from input uncertainty (e.g. air pollution emissions inventories), or limitations of the transport and chemistry mechanisms within the model (Brynjarsdottir and O’Hagan, 2014). The current interest of this study focuses on a better estimate of mean ozone exposure. Explicit quantification of different sources of model uncertainty and incorporation of this information into the data fusion process presents another level of complexity that cannot be tackled until model uncertainties are better characterized. Young et al. 2018 provide a current overview of chemistry-climate modelling and discuss the challenges of improving models in light of so many uncertainties.”

2.2 For example, creating an ensemble of weights (and therefore and ensemble of fused data sets) could be used to explore the impact of poor observational sampling on the fused data set compared to the multi-model mean.

We expanded the discussion on the differences between our fused product (also model weighting product) and the multi-model mean at the end of Section 3:

“When interpreting the fused product the reader should consider the following: (1) For a region with an extensive monitoring network, such as the USA, a detailed bias correction can be achieved. We can utilize the observations to accurately reflect many local features (i.e., sub-grid variations) as shown in the ozone pollution hot-spots of southern California and Mexico City. However it should be noted that this improvement is due to bias correction, instead of model weighting; (2) For regions with large observational gaps, such as South America, Africa or Russia, the spatial difference between the fused product and the multi-model mean is rather featureless, because the model weighting can only adjust the overall regional mean according to a few monitoring sites, and cannot address the local variations. Filling large data gaps with the intermediate multi-model composite can indeed avoid the influence of preferential sampling (Diggle et al., 2010; Shaddick and Zidek, 2014), but it is still subject to a high uncertainty due to lack of data.”

3. In order to compare both the interpolated observations and each models, and the multi-model mean with the fused dataset, I suggest to also plot the empirical variograms, to quantify the differences in the spatial structure.

Thanks, the variogram is indeed a useful tool to summarize the spatial structure. We added a discussion about variogram in the end of Section 3.3

“The fused product can be evaluated in terms of spatial correlation using the variogram which assumes that spatial correlation is not a function of absolute location, but only a function of distance (i.e., stationarity). Since spatial variability and continuity from the models are the result of geophysical processes represented by mathematical equations, the variogram must be customized for each field. In addition, the extremely large size of the model output prohibits us from carrying out a standard empirical variogram analysis, which requires calculating the variance of the difference between all pair-wise grid cells.

Nevertheless, we provide examples of omnidirectional variograms for the spatial field in North America from each model and product in supplementary Fig. S-5. The standard variogram analysis focuses on the following three parameters: (1) the nugget (variance at zero distance, which represents a sub-grid variation), which is similar for all cases; (2) the sill (total variance of a field), where the variogram value reaches a maximum and levels off The result is very similar for G5NR-Chem, GEOSCCM and GFDL-AM3, while CHASR and MRI-ESM show a larger variance in the spatial field. The reason is that the latter two models produce low ozone in the high latitude region over Canada (see supplementary Fig. S-1), but the former three models simulate relatively higher ozone in the same region, and this difference is reflected by the total variance; (3) the range (a distance where the sill is reached, and beyond that there is no longer spatial correlation): the variogram peak is about 35-40 degrees for the models. Note that a continuously increasing variogram indicates the evidence of non-stationarity in the field, which is the case for SPDE, an issue that we have accounted for. Even though North America has one of the most extensive monitoring networks in the world, some of the remote areas (mostly in Canada) are mainly described by the model output in the final fused product. Therefore the variogram of the fused product is likely adjusted toward the remote areas of Canada as simulated by G5NR-Chem, which provided the largest weighting in North America).”

4. Line 27, page 6: cite the R core development team.
A citation was added to the code and data availability section.

Anonymous Referee #2

We thank the reviewer for valuable comments on our manuscript. The comments from the reviewers are below in bold font and we make a response accordingly.

General comments. Overall quality

The article proposes a method for combining measurements from 6 different global models with the aim of generating an improved estimation of the global surface ozone when compared to the estimation obtained by the simple average of these 6 different global models. Hence, this article proposes a method for estimating the weight factor to give to each global model within a weighted average of global models available, and also proposes a method for fusing this result with kriging estimates depending of closeness of locations to monitoring networks. The latter results in an estimated global surface ozone which is a combination between interpolation-based kriging for areas near monitoring networks, and the proposed weighted average for areas far from monitoring networks. Notwithstanding, results from this article and the final surface for global ozone is estimated by a smoothing splines approach which is applied to the estimation either of the composite model or its fused version. Consequently, the smoothing splines step is the key for the method presented, however it is not explained in the article and authors only dedicate two to three lines to comment about its use to avoid discontinuities in the joints between continental regions.

Thank you for pointing out this issue. We indeed need to clarify that the use of the smoothing splines is not a key method for producing the final fused product, and is only a minor step that we employ to smooth the transition between three regions with sharp discontinuities. This smoothing is only conducted over a 5-degree distance along the boundaries between 2 regions at the 3 locations in the world with the largest discontinuities (see below Figure 1), leaving the rest of the regions unaffected. This procedure is now more clearly described in Section 2.2 (step 2):

“This smoothing is carried out using a low rank Gaussian process by the default penalized least square from the function ``gam'' in the R package mgcv (Kammann and Wand, 2003; Wood, 2017), following the examples of Wood and Augustin (2002).The purpose is to merely avoid a sharp and unrealistic (geometric) transition between three regions and to efficiently smooth out the discontinuity, performed in a regular spaced grid only around the geometric boundary. Any region away from the geometric boundary will not be affected by this smoothing, which should be considered as a blending of multiple models without any attempt of bias correction (see supplementary Fig. S-2).”

We added Figure 1 to the Supplement (Fig. S-2) to illustrate the smoothing procedure, using an expanded color scale to highlight the impact of the spline smoothing on the regional discontinuities. We only apply the smoothing to 3 regions: one horizontal discontinuity between Russia and East/South Asia, one vertical discontinuity between East and South Asia, and one vertical discontinuity between South Asia and Africa. The discontinuities along the rest of the boundaries between regions were minor and therefore we do not make any further adjustment.

The spatial structure produced from the weighting is not supposed to be create a discontinuity, because the output is smooth. However, a straight line discontinuity is an artifact from our regionalized module in which the models are evaluated for separate regions of the world; the spline smoothing corrects this artifact.

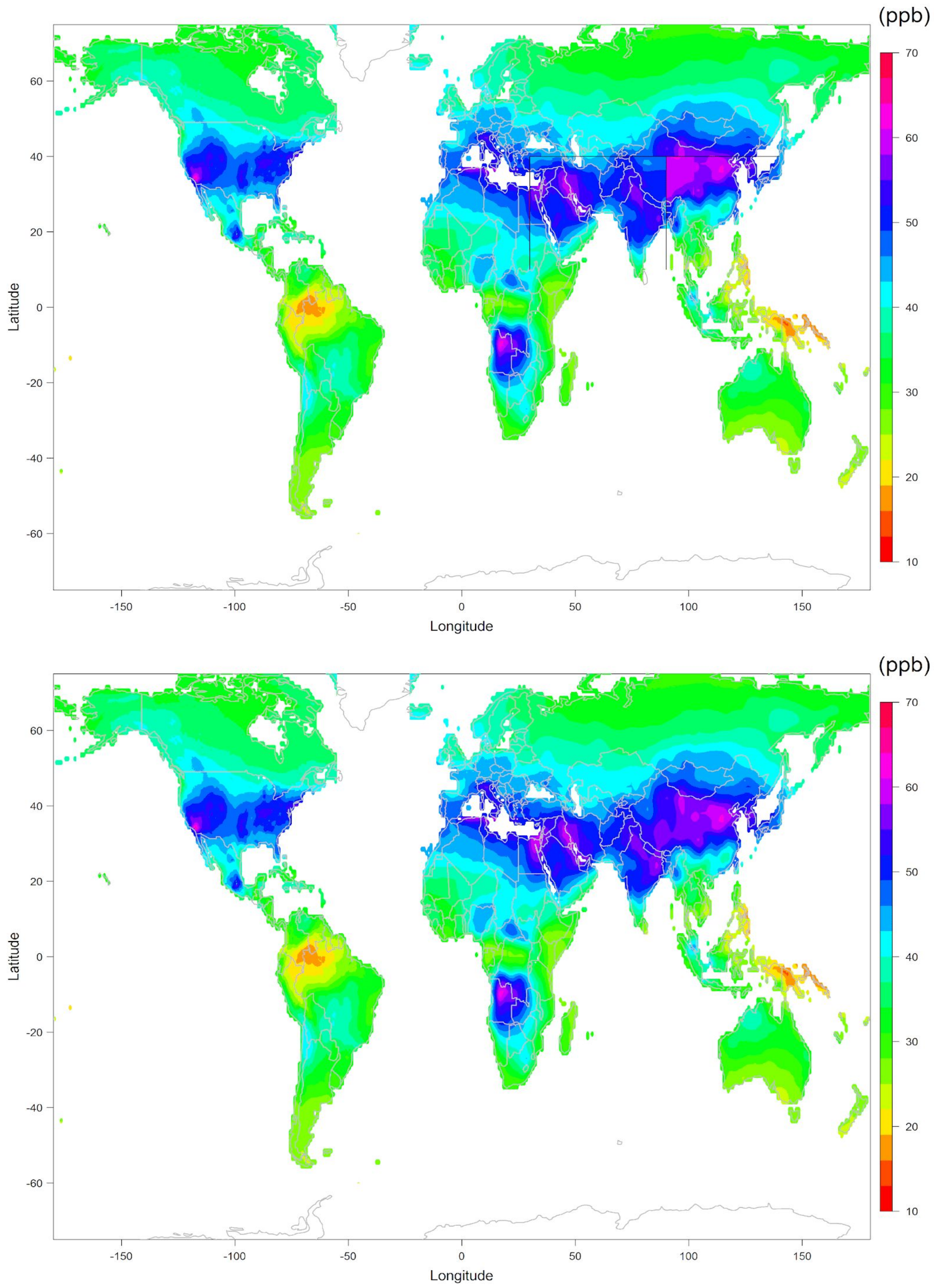


Figure 1: Strong ozone discontinuities, or artifacts, were present along the boundaries between world regions, especially in western China, before a spline smoothing was employed.

The authors are trying to address three different problems in geostatistics. First, irregularly spaced sources of information or when the coordinates of the locations from different sources do not match. Second, the lack of information due to observations sparsely distributed or missing locations or almost no information in certain regions. And third, to obtain a better estimate of a surface or compare one estimate with others.

The first problem is more related with interpolation and this is explained well in sections 2.1 and 2.2 of the article. The second problem is being addressed by the use of global ozone models to obtain a better guess of the non-observed ozone in certain locations. Here the authors propose the composite mean between the global models and its fused version with the interpolation depending on closeness of monitoring network stations which in practice is working as a method for "imputation" of ozone in non-observed locations. The description of the method is mostly well explained (although it is missing important details which I describe in the next section of this report) but the method is not a solution for this problem in areas like Africa or South America where there is not enough information and this is not solved by the composite nor fused method, but by having more measurements. This should be stated clearly in the article. It is difficult to believe that the weight coefficients estimated for Africa or South America would be good estimates given the little sample size available. Nevertheless, the authors could have taken advantage of the dense data available for North America, Europe and East Asia to perform cross-validation and then using the data from South America, Africa and Australia as validation sets of data. This would have provided a rough idea of the quality of the composite estimates to perform the imputation of the ozone level on areas with sparsely distributed observations.

We modified the text associated with large data gaps in Section 2.2:

“Above land, large observational gaps are present across Africa, the Middle East, South America, and South and Southeast Asia, where the spatial interpolation is generally too uncertain to yield a reliable surface ozone approximation. The ozone estimates in these regions must come from either models or distant observations, neither of which is ideal to solve this issue. As a compromise strategy we fill these gaps with a weighted model product evaluated by the interpolated ozone observations.”

Following the recommendations of the referee we have expanded this discussion at the end of Section 3:

“The advantage of our fused surface ozone product over the simple multi-model mean can be clearly seen in Figure 8. When interpreting the fused product the reader should consider the following: (1) For a region with an extensive monitoring network, such as the USA, a detailed bias correction can be achieved. We can utilize the observations to accurately reflect many local features (i.e., sub-grid variations) as shown in the ozone pollution hot-spots of southern California and Mexico City. However it should be noted that this improvement is due to bias correction, instead of model weighting; (2) For regions with large observational gaps, such as

South America, Africa or Russia, the spatial difference between the fused product and the multi-model mean is rather featureless, because the model weighting can only adjust the overall regional mean according to a few monitoring sites, and cannot address the local variations. Filling large data gaps with the intermediate multi-model composite can indeed avoid the influence of preferential sampling (Diggle et al., 2010; Shaddick and Zidek, 2014), but it is still subject to a high uncertainty due to lack of data."

The cross validation technique is indeed a common criterion for assessing the spatial fits, and we used a simple leave one out (LOO) cross validation for assessing our model (the GCV score in table A1). However this score represents an overall LOO error, and doesn't allow for an observation in sparsely sampled region, such as Africa, to receive a lower fitted error than any other observation (and it should not because it would be a conceptual prejudice). We prefer to avoid this type of analysis as we cannot explicitly quantify the representativeness of every single site.

The third problem is poorly or not explained in the article. As presented, the article gives the impression that the main modelling product is the composite mean or its fused version which is confusing since the results are based on a smoothed version and this is not explained in the article. One reader could think that in equation (1) the yˆ(sg)'s are the "imputed" observations based on the interpolation technique while the composite mean is the proposed model for the ozone level. However, given that later in the article it is expressed that the results are obtained using smoothing splines over the fitted composite mean surface or its fused version, other readers can interpret that actually the fitted composite or fused mean are the imputed observations of the ozone level and the proposed model is the smoothing spline. This is extremely confusing and the article is poorly explained in all this part.

Thanks for the commentary. We hope we have clarified this concern over the smoothing splines, as described above. The model composite is indeed made from the models, while the smoothing splines only play a minor role for the purpose of removing the straight line discontinuities in Asia and Africa. We further clarified this point in Section 2.2 (step 2):

"It should also be noted that the INLA-SPDE technique in step 1 is applied to the observations, while the smoothing spline is only applied to the boundaries between regions of the model composite, not directly involving any observations."

Specific comments. Individual scientific questions/issues

(a) There are important issues which are not addressed by the authors. What is the real role of the smoothing spline applied to the fussed estimation as described in page 8 lines 10-15 and the supplementary material Figure S-2? As the article is presented, this step seems to have a minor role for their proposed method, however it is a key step and the authors did not explain this in detail. In the abstract and along the presentation until

section 2.2, the authors' product of this work is a method which relies on a fusion between a weighted average of the 6 global models and the interpolation/kriging step. Nevertheless, from the results it can be deduced that the final product of this work is actually the smoothing spline fit to the surface obtained either by the composite method or its fused version. Therefore, the results presented in Figure 8 (and therefore all related results) rather than being the surface obtained by the multi-model composite and multi-model composite plus bias correction, are respectively the smoothing spline fit to the surface obtained by the multi-model composite and the smoothing spline fit to the surface obtained by the multi-model composite plus bias correction. This must be clearly stated.

To illustrate the limited impact of the spline smoothing on just three regional boundaries we have included figure S-2 in the Supplement.

(b) What is the interpretation of the parameters in every model discussed? For example:
-What is the practical interpretation of the parameter αr in equation (1)? It is related to the general mean over the r-th continental region.
-What is the practical interpretation of the parameter βrk in equation (1)? The cell-by-cell average model corresponds to assume βrk = 1/6, thus the same weight is given to each model on each continental region. Then on the composite model βrk can have a meaningful interpretation but the authors do not comment on this.
-What does it mean if βrk = 0 (i.e. with respect to the cell-by-cell mean model)? Moreover, what does it mean if βrk 6= 0? How to interpret if βrk < 0? How to interpret if βrk > 0? This is important and connects the cell-by-cell average with the composite mean model proposed. It tells us whether the composite model offers or not a better representation than the average model.

Since the interpolated observations and models use the same ozone metric with the same units, it makes sense that we restrict the coefficient of the covariate to a range between [0, 1] and summed to 1:

- From a regression point of view, if we only include beta (weight) without a constant, the residuals will have a biased mean, so the alpha term will force the overall residuals to have a mean value of zero.
- Any positive value of beta should be seen as a significant component of the model composite.
- If beta is zero, it means this particular model makes no contribution to the model composite.

The coefficient is not permitted to have a negative value since a negative value doesn't have a physical meaning.

We modified the interpretation in Section 2.2 (step 2):

"Note that since the interpolated observations and models use the same ozone metric with the same units, we thus constrain the weights to be positive and sum to 1 for better physical interpretability, such that the most accurate models receive the higher weight. A constant offset αr is included to guarantee that the residuals from this optimization have a zero mean."

(c) Regarding comment (b), some statistical summaries are not presented in Table 2. For example, what is the significance of each weight/coefficient for the global models and their standard errors or confidence limits? Note that the proposed quadratic programming idea can be seen as a multiple linear regression within each continental region where the yˆ(sg)'s are seen as the (imputed) observations, the global ozone models ηr1(sg), . . . , ηr6(sg) are seen as predictors or covariates, and the errors are assumed uncorrelated with constant variance. From this approach the authors can obtain variability estimates for the weight coefficients and test their significance.
There is no standard or consensus methodology for combining models based on the uncertainty (i.e., standard errors), and we are unable to adjust the weights based on the standard errors. For example, no matter what value of standard error is associated with a 0 coefficient, it will still have a 0 weight, which doesn't allow us to properly interpret the variability. This is why we restrict the coefficients to a range between [0, 1] and summed to 1; this arrangement forces the coefficient of an insignificant predictor to be 0, and any positive coefficient should be seen as a significant contribution to the model composite. We added a discussion on the difficulty of combining models based on the uncertainties to the Section 2.2:

"We adopted a regression weighting approach that only accounts for the mean spatial fields of the interpolated ozone and model output, rather than the underlying associated uncertainty. We take this approach due to the prohibitive size of high resolution output (over 1 million output points for each model), but also due to the lack of a thorough investigation regarding the ideal method for combining models based on different sources of uncertainty. For example, the interpolation uncertainty can be quantified easily through the posterior distribution and considered to be related to measurement error (small scale) or sparse sampling across a region (large scale), however, model uncertainty is a different concept altogether that could result from input uncertainty (e.g. air pollution emissions inventories), or limitations of the transport and chemistry mechanisms within the model (Brynjarsdottir and O'Hagan, 2014). The current interest of this study focuses on a better estimate of mean ozone exposure. Explicit quantification of different sources of model uncertainty and incorporation of this information into the data fusion process presents another level of complexity that cannot be tackled until model uncertainties are better characterized. Young et al. 2018 provide a current overview of chemistry-climate modelling and discuss the challenges of improving models in light of so many uncertainties."

(d) The authors comment that their composite and fused composite method is better than the simple average (or cell-by-cell average) method. It would be helpful if the authors presented p-values for a test comparing these two hypotheses.
To the best of our knowledge the use of a hypothesis test or p-value for comparing spatial model fits (or climate model performance) is not an ideal approach and not discussed in the literature. This is largely because the kriging procedure (or Gaussian process) is the result of machine learning, so there is no corresponding "hypothesis testing" concept for the Gaussian process. Rather, computer scientists tune the parameters to yield the best output.

The most common practice to measure the performance of a model is directly comparing the root mean square error (RMSE, as shown in tables 2 and 3) between observations and output, and quantifying the percentage of improvement. The report of the physical quantity, such as RMSE shown in the same unit as the ozone metric, should be more meaningful than potentially misleading p-values.

(e) The authors do not mention the assumption for the mean nor covariance of the smoothing splines model, nor give any details about which type of splines they used (tensor products, thin-plate splines, regression splines, etc.). Did you use penalties? The authors only refer to mgcv R package (Wood, 2017) in line 20 of page 13 and we need to see the code to see what they did, however they should also explain their method, procedure and assumptions in the article. It is the most important modelling they are doing and their results depend on this smoothing splines step.

As described above, the spline smoothing was just a very minor component of our method, only used to smooth 3 geographical discontinuities. We used a particularly simple form of the Matern covariance function suggested by Kammann and Wand (2003), and we added the details to Section 2.2.

The main command to perform this smoothing is given by (a complete code can be found in supplementary material):

```
mod = gam(composite ~ s(lon,lat, bs="gp", k=180), data=sm, method="REML",
na.action='na.omit')
```

Since the removal of a straight line discontinuity is the only concern, any spline model should achieve this goal as long as it can handle the high resolution output. We chose this Matern spline merely because it is simple and efficient for high resolution output.

(f) We can see three different steps in this method. The initial interpolation using INLA, the determination of the weights, and the final smoothing splines using mgcv. INLA and the composite are imputing the ozone measurement on unobserved locations, and the gam function of mgcv package is performing the fit using smoothing splines. In practice, INLA and the smoothing splines are performing the same procedure: interpolation. The only difference is that INLA is based on a triangulation and finite element approach to find a solution. Besides, in both cases the authors are assuming a Matérn covariance function. Therefore, in practice they are fitting an interpolation model to the data (using INLA), and then fitting pretty much the same interpolation model (but using mgcv package) to the previous fit obtained by INLA. Thus, the INLA interpolation is "smoothing" the variation (Figure 3a, page 28), and then an additional smoothing using gam function is being performed (Figure 8a and 8b, page 33, and Figure S-4 in supplementary material). These two fits seem very similar and differences between them can be (visually) attributed mainly to the "variation" generated by the composite mean fit (Figure S-2).

Yes the INLA and smoothing spline are indeed performing the same procedure, but as described above the spline smoothing is only used under very limited circumstances to smooth 3 regional boundary discontinuities. We also added a note in Section 2.2 (step 2):

"It should be noted that the INLA-SPDE technique in step 1 is applied to the observations, while the smoothing spline is only applied to the boundaries between regions of the model composite, not directly involving any observations."

-The first INLA smoothing imputes the ozone at unobserved locations but the resulting "smoothed" process has smaller variation than what we could expect from the original spatial process. Why did the authors not use resampling methods for the imputation step (either before applying the INLA interpolation and/or before applying the smoothing splines fit)? This would have allowed them to keep some spatial variability on the imputed spatial process, and also evaluate assumptions regarding this spatial variability model.
-Examples of how to implement resampling methods can be found in Liang et al. (2013), and in a more practical presentation by Muñoz et al. (2010). Other approaches based on Expectation-Maximization (EM) algorithm are presented in Schneider (2001).

When comparing Figures 1 and 2(a), the INLA interpolation can reproduce the hot spots in East China, Mexico City and LA, but it indeed missed the highest ozone in the Beijing Metropolitan area. This result is related to the degree of smoothing that we allow for the spatial field. The success of reproducing ozone in East China, Mexico City and LA is due to these locations having multiple grids with high ozone observations. However, there is only one grid with high ozone observed in Beijing, and half of the grid points around that spot do not observe high ozone, therefore the smoothing of the spatial field missed this single hot spot. Reducing the degree of smoothing in order to capture the ozone hot spot in Beijing Metropolitan would introduce more noise and unrealistic peaks in other regions, and also increase the computational burden.

We choose the INLA-SPDE technique for interpolation merely because it can incorporate the non-stationary component in an easy way (illustrated in the appendix). No matter the details of the INLA-SPDE, resampling method, or other approaches mentioned in Section 2.2 (covariance tapering, low rank, spectral representation, likelihood approximation…), they are all special cases of a more general Gaussian process designed to alleviate the large sample size (n) problem. Parts of these approaches are only proven to be efficient on regional or national scales, and they are not necessarily adequate on the global scale. For example in the differential manifold, a covariance model is positive definite on a plane, it is not guarantee that will also valid in a sphere (Gneiting, 2013).

The idea of resampling is similar to the leave-n-out validation, it keeps iteratively removing a portion of data and re-fitting the model until the error is minimized. So it is an algorithm to improve the fit to the data (we cannot leave the data out without an actual observation).

However, the similar but much simpler leave-1-out validation is used for evaluating the interpolation performance (the GCV score in table A1). We added the citations accordingly in the manuscript, but Munoz et al. (2010) focused on the discussion of type and mechanism of missing data, which is a bit far from our topic.

(g) Regarding the previous comment (f), the authors did not present results about the estimation of the Matérn semivariogram's parameters for the INLA and smoothing splines step. Given that they are performing a "pre-smoothing" of the variation using INLA, it would be expected that the Matérn semivariogram in the smoothing splines step would be modelling a significantly lower amount of spatial variation which might result in almost uncorrelated errors (except in areas where there are peaks or throughs in the process). Is that a good representation or assumption for the global ozone process?
The INLA package provides a more general class of model that can specify a spatially varying nugget and sill, which would be more flexible than the variogram approach that assumes a fixed nugget and sill over the whole spatial field. From a series expansion of spherical harmonics in Eq A1, we used several basis functions to select the best statistical model in table A1; for further details please refer to the appendix. We also added a discussion about variograms to the end of Section 3.3

"The fused product can be evaluated in terms of spatial correlation using the variogram which assumes that spatial correlation is not a function of absolute location, but only a function of distance (i.e., stationarity). Since spatial variability and continuity from the models are the result of geophysical processes represented by mathematical equations, the variogram must be customized for each field. In addition, the extremely large size of the model output prohibits us from carrying out a standard empirical variogram analysis, which requires calculating the variance of the difference between all pair-wise grid cells.

Nevertheless, we provide examples of omnidirectional variograms for the spatial field in North America from each model and product in supplementary Fig. S-5. The standard variogram analysis focuses on the following three parameters: (1) the nugget (variance at zero distance, which represents a sub-grid variation), which is similar for all cases; (2) the sill (total variance of a field), where the variogram value reaches a maximum and levels off The result is very similar for G5NR-Chem, GEOSCCM and GFDL-AM3, while CHASR and MRI-ESM show a larger variance in the spatial field. The reason is that the latter two models produce low ozone in the high latitude region over Canada (see supplementary Fig. S-1), but the former three models simulate relatively higher ozone in the same region, and this difference is reflected by the total variance; (3) the range (a distance where the sill is reached, and beyond that there is no longer spatial correlation): the variogram peak is about 35-40 degrees for the models. Note that a continuously increasing variogram indicates the evidence of non-stationarity in the field, which is the case for SPDE, an issue that we have accounted for. Even though North America has one of the most extensive monitoring networks in the world, some of the remote areas (mostly in Canada) are mainly described by the model output in the final fused product. Therefore the *variogram of the fused product is likely adjusted toward the remote areas of Canada as simulated by G5NR-Chem, which provided the largest weighting in North America).”*

The multi-model mean and composite (Figures 3(a) and 7(a)) show a lower spatial variation, which is unrelated to INLA or the Matern function, since they are completely made/weighted from the model output. Based on the comparison of the similarity between Figures 1 and 2(a), we see that the interpolation reproduces many local features, a result that we view as successful.

(h) Regarding the modelling part, as presented the article overlooks the real role that the smoothing splines step (page 8 lines 8-15 and Figure S-2 in supplementary material) is playing in the resulting global ozone surface estimated. What is the main modelling technique of their method: the composite method with/without bias correction, or the smoothing splines in question? This is a key issue which cannot be disregarded.
As described above, the smoothing splines are only a minor component of the overall analysis. Our final fused product is the result of three important steps, all of which are necessary. We modified the following paragraph in the Introduction that summarizes the key steps in this process:

“This paper presents a new statistical approach (M3Fusion) for combining surface ozone output from multiple atmospheric chemistry models with all available surface ozone observations to produce a global surface ozone distribution with greater accuracy than the multi-model ensemble mean. As described in greater detail below, this fused surface ozone product is constructed in three steps: 1) Ozone observations from all available surface ozone monitoring sites around the world are spatially interpolated to a smooth global field; 2) For each of 8 continental regions of the world 6 global atmospheric chemistry models are evaluated against the interpolated observed ozone field by a quadratic programming optimization, with the most accurate models receiving the highest weight. A locally confined spline interpolation is used at the regional boundaries to avoid unphysical step changes; 3) finally, the global ozone field derived from the polynomial equation is bias corrected, but only within a limited distance from available observations. The final product is based on the annual maximum of the 6-month running mean of the monthly average daily maximum 8-hour average mixing ratios (DMA8), a metric that can be used to estimate human mortality due to long-term ozone exposure (Turner et al., 2016; Malley et al., 2017; Seltzer et al., 201; Shindell et al.,2018).”

(i) At moments, it is not clear whether the key goal of the article is to propose a novel method to estimate the weights to give to each global ozone model, or to propose an estimated global ozone surface (which is indeed being obtained based on the smoothing splines step). I suggest this is clarified.
The goal is to create a best estimate surface for the global ozone concentration, and this is done through a three-step process (which we clarified in the paragraph as above), and we also added a note in Conclusion that this method can be used for different applications:

“The application of our methodology focuses on, but is not limited to, a particular ozone metric relevant for quantifying the impact of long-term ozone exposure on human health. We expect that this framework could also be applied to other ozone metrics relevant to crop production or natural vegetation (Lefohn et al., 2018; Mills et al., 2018), or any other trace gas provided adequate in situ observations are available for model evaluation.”

(j) The authors mention that the success of their composite mean obtained via the quadratic programming approach depends on the existence of a global ozone model which reproduces correctly the correct curvature on the process (line 30 page 7 – line 5 page 8). This is not necessarily true since equation (1) is only selecting αr, βrk based on a least squares criterion and no regularization conditions on the solution are being specified, i.e. a curvature penalty. Besides, the curvature of a surface is defined throughout the two-dimensional space in question (the map), and requires the existence of first and second derivatives of the surface. None of these conditions are being established in the article, so that the weights of the composite mean are best only in terms of the "squared distance" between the (imputed) observations and the composite mean at the regular grid of locations being used, not throughout the continuous two-dimensional space.

Thanks for the suggestion, we indeed have not discussed about spatial curvature and first/second derivatives of the surface. We thus removed this paragraph and added a clarification in p7:

“The weights are optimized in terms of the squared distance between the interpolated ozone and multi-model output. A different criterion of optimization, such as mean absolute error, can be established accordingly.”

Additional (tentative) references

Liang, Faming, Yichen Cheng, Qifan Song, Jincheol Park, and Ping Yang. "A resampling-based stochastic approximation method for analysis of large geostatistical data." Journal of the American Statistical Association 108, no. 501 (2013): 325-339.

Muñoz, Breda, Virginia M. Lesser, and Ruben A. Smith. "Applying Multiple Imputation with Geostatistical Models to Account for Item Nonresponse in Environmental Data." Journal of Modern Applied Statistical Methods 9, no. 1 (2010): 27.

Schneider, Tapio. "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values." Journal of climate 14, no. 5 (2001): 853-871.

Reference:

Brynjarsdóttir, J. and O’Hagan, A.: Learning about physical parameters: The importance of model discrepancy, Inverse Problems, 30, 114 007, 2014.

Diggle, P. J., Menezes, R., and Su, T.-I.: Geostatistical inference under preferential sampling, J. Roy. Stat. Soc. C, 59, 191–232, 2010.

Gneiting, T. (2013) Strictly and non-strictly positive definite functions on spheres. Bernoulli, 19, 1327–1349.

Kammann, E. and Wand, M. P.: Geoadditive models, J. Roy. Stat. Soc. C, 52, 1–18, 2003.

Lefohn, A. S., Malley, C. S., Smith, L., Wells, B., Hazucha, M., Simon, H., Naik, V., Mills, G., Schultz, M. G., Paoletti, E., De Marco, A., Xu, X., Zhang, L., Wang, T., Neufeld, H. S., Musselman, R. C., Tarasick, D., Brauer, M., Feng, Z., Tang, H., Kobayashi, K., Sicard, P., Solberg, S., and Gerosa, G.: Tropospheric ozone assessment report: Global ozone metrics for climate change, human health, and crop/ecosystem research, Elem. Sci. Anth., 6, 2018.

Mills, G., et al. (2018), Tropospheric Ozone Assessment Report: Present-day tropospheric ozone distribution and trends relevant to vegetation, Elem. Sci. Anth., 6(1):47.

Shaddick, G., and Zidek. J.V.: A case study in preferential sampling: Long term monitoring of air pollution in the UK. Spatial Statistics 9 (2014): 51-65.

Wood, S. N. and Augustin, N. H.: GAMs with integrated model selection using penalized regression splines and applications to environmental modelling, Ecol. Model., 157, 157–177, 2002.

A new method (M^{3}Fusion-v1) for combining observations and multiple model output for an improved estimate of the global surface ozone distribution

Kai-Lan Chang[1, 2, 3], Owen R. Cooper[2, 3], J. Jason West[4], Marc L. Serre[4], Martin G. Schultz[5], Meiyun Lin[6, 7], Virginie Marécal[8], Béatrice Josse[8], Makoto Deushi[9], Kengo Sudo[10, 11], Junhua Liu[12, 13], and Christoph A. Keller[12, 13, 14]

[1]National Research Council Fellow
[2]NOAA Earth System Research Laboratory, Boulder, CO, USA
[3]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA
[4]Department of Environmental Sciences & Engineering, University of North Carolina, Chapel Hill, NC, USA
[5]Jülich Supercomputing Centre (JSC), Forschungszentrum Jülich, Jülich, Germany
[6]NOAA Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA
[7]Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ, USA
[8]Météo-France, Centre National de Recherches Météorologiques, Toulouse, France
[9]Meteorological Research Institute (MRI), Tsukuba, Japan
[10]Graduate School of Environmental Studies, Nagoya University, Nagoya, Japan
[11]Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokosuka, Japan
[12]NASA Goddard Space Flight Center, Greenbelt, MD, USA
[13]Universities Space Research Association, Columbia, MD, USA
[14]John A. Paulson School of Engineering and Applied Science, Harvard University, Cambridge, MA, USA

Correspondence: Kai-Lan Chang (kai-lan.chang@noaa.gov)

Abstract. We have developed a new statistical approach (M^{3}Fusion) for combining surface ozone observations from thousands of monitoring sites around the world with the output from multiple atmospheric chemistry models to produce a global surface ozone distribution with greater accuracy than can be provided by any individual model. The ozone observations from 4766 monitoring sites were provided by the Tropospheric Ozone Assessment Report (TOAR) surface ozone database which contains the world's largest collection of surface ozone metrics. Output from six models was provided by the participants of the Chemistry-Climate Model Initiative (CCMI) and NASA's Global Modeling and Assimilation Office (GMAO). We analyze the 6-month maximum of the maximum daily 8-hour average ozone value (DMA8) for relevance to ozone health impacts. We interpolate the irregularly-spaced observations onto a fine resolution grid by using integrated nested Laplace approximations, and compare the ozone field to each model in each world region. This method allows us to produce a global surface ozone field based on TOAR observations, which we then use to select the combination of global models with the greatest skill in each of 8 world regions; models with greater skill in a particular region are given higher weight. This blended model product is bias-corrected within two degrees of observation locations to produce the final fused surface ozone product. We show that our fused product has an improved mean squared error compared to the simple multi-model ensemble mean, which is biased high in most regions of the world.

1 Introduction

Tropospheric ozone is a pollutant detrimental to human health and has been associated with a range of adverse cardiovascular and respiratory health effects due to short-term and long-term exposure (World Health Organization, 2005; Jerrett et al., 2009; US Environmental Protection Agency, 2013; GBD, 2015; Turner et al., 2016; Cohen et al., 2017). Assessing the human health impacts of ozone on the global scale requires accurate exposure estimates at any given inhabited location (Shaddick et al., 2018). Due to the limited availability of surface ozone observations in many regions of the world (Fleming et al., 2018), global atmospheric chemistry models are required to calculate surface ozone exposure. Despite continual development and improvement, global models struggle in their ability to accurately simulate ozone in all regions of the world (Young et al., 2018). The ability to accurately simulate observed ozone at a particular location also varies between models, as demonstrated by several multi-model comparisons (Stevenson et al., 2006; Young et al., 2013; Cooper et al., 2014).

A useful endeavor for producing an accurate representation of the global surface ozone distribution is to combine the output from many models in a way that takes advantage of the strengths of each model and minimizes the weaknesses. Such efforts have already been made for both climate and chemistry climate models. For example, multi-model output has been combined using a parametric approach, either by assigning an equal or optimum weight to each model (Stevenson et al., 2006; He and Xiu, 2016; Braverman et al., 2017), or by tuning the initial conditions under different scenarios or parameterizations (Cariolle and Teyssèdre, 2007; Wu et al., 2008; Young et al., 2013). These approaches often assume that individual model biases will at least partly cancel by averaging or weighting, according to certain measures of predictive performance. Thus the combined model product is likely to be more accurate than a single model prediction, as has been shown for multi-model combinations of past or present day climate (Buser et al., 2009; Knutti et al., 2010; Weigel et al., 2010; Chandler, 2013).

For the case of simply averaging the output from multiple climate models, most studies either explicitly or implicitly assume that every model is independent and is a random sample from a distribution, with the true climate as its unbiased mean. This implies that the average of a set of models converges to the true climate as more and more models are added. This multi-model ensemble often outperforms any single model in terms of the predictive capability. Undeniably, when one has several dozen or hundreds of possible ensemble members, the most straightforward and efficient approach is to simply take the ensemble average, ignoring the impact of potentially erroneous outlier ensemble members. From a statistical point of view, one might argue that ruling out potentially erroneous ensemble members prior to conducting the ensemble mean would yield an even better result, especially if the overall number of ensemble members is small.

Combining model ensembles using a method more sophisticated than the simple average is a challenge because a meaningful model evaluation can rarely be condensed into a single metric, and there is no technique that can explicitly quantify the degree of similarity (i.e both accuracy and precision) between two different spatial fields (Hyde et al., 2018). Indeed, Stainforth et al. (2007) concluded that any attempt to assign weights is, in principle, inappropriate. With a lack of appropriate criteria, the model weighting approach has not become a standard alternative to the ensemble average. Accordingly, there is presently no objective criterion for combining surface ozone estimates from a model ensemble to produce a surface ozone product with improved accuracy beyond that of any ensemble member or the simple ensemble mean. The absence of such a methodology is the motivation for this paper.

This paper presents a new statistical approach (M^3Fusion) for combining surface ozone output from multiple atmospheric chemistry models with all available surface ozone observations to produce a global surface ozone distribution with greater accuracy than the multi-model ensemble mean. As described in greater detail below, this fused surface ozone product is constructed in three steps: 1) Ozone observations from all available surface ozone monitoring sites around the world are spatially interpolated to a smooth global field; 2) For each of 8 continental regions of the world 6 global atmospheric chemistry models are evaluated against the interpolated observed ozone field by a quadratic programming optimization, with the most accurate models receiving the highest weight. A locally confined spline interpolation is used at the regional boundaries to avoid unphysical step changes ; 3) finally, the global ozone field derived from the polynomial equation is bias corrected, but only within a limited distance from available observations. The final product is based on the annual maximum of the 6-month running mean of the monthly average daily maximum 8-hour average mixing ratios (DMA8), a metric that can be used to estimate human mortality due to long-term ozone exposure (Turner et al., 2016; Malley et al., 2017; Seltzer et al., 2018; Shindell et al., 2018).

Past estimates of global mortality due to long-term ozone exposure have relied on surface ozone fields produced by global atmospheric chemistry models due to the limited coverage of the global ozone monitoring network (Anenberg et al., 2010; Brauer et al., 2012, 2015; Malley et al., 2017). The fused surface ozone product is a blend of global surface ozone observations and model output that has been adjusted according to the observations. This particular product will be available for future estimates of global human mortality due to long-term ozone exposure (e.g. Global Burden of Disease (Brauer et al., 2012, 2015)). Furthermore, the methodology can be applied to a range of ozone metrics for quantifying the impacts of ozone on human health, or vegetation, and it can also be applied to $PM_{2.5}$, CO_2, or any other trace gas.

Section 2 provides details of the data sources and fusion process, including the techniques to register all data sources onto a common grid, and the statistical model used to minimize the difference between interpolated observations and the multi-model combination. In Section 3 the results of employing these techniques are presented, including the mapping accuracy, evaluation of regional model performance and the final multi-model bias correction. The paper concludes with a summary and discussion in Section 4.

2 Data and Method

2.1 Observations and model output

1. *Tropospheric Ozone Assessment Report (TOAR) Surface Ozone Database*: In this analysis, surface ozone observations are used to evaluate the performance of 6 global atmospheric chemistry models and to also bias-correct the multi-model surface ozone product. TOAR has produced the world's largest database of surface ozone metrics based on hourly observations at over 9000 sites around the globe (Schultz et al., 2017, ozone metrics available for download at: https://doi.org/10.1594/PANGAEA.876108). Spatial coverage is high in North America, Europe, South Korea and Japan, but much lower across the rest of the world with very low data availability across Africa, the Middle East, Russia and India.

In addition to data sparseness, other challenges, such as data inhomogeneity in time and the irregular spatial distribution of stations (Chang et al., 2017), make the comparison between model output and observations difficult without serious statistical modeling. While satellite retrievals have been utilized by previous works for quantifying the health impacts of $PM_{2.5}$ (Brauer et al., 2012, 2015), satellite retrievals of tropospheric ozone have limited sensitivity near the surface and are inadequate for this analysis (Gaudel et al., 2018).

TOAR has gathered ozone observations through 2014 at most sites, and has chosen 2008-2014 as a “present-day” window for more rigorous analysis. The purposes of the multi-year average are to reduce the effects of ozone interannual variability, which is largely driven by changes in meteorological conditions (Strode et al., 2015), and to increase the number of available sites than if we used a single year. In this analysis we focus on the annual maximum of the 6-month running mean of the maximum daily 8-hour average (DMA8) at every site in the TOAR database. Specifically, the metric was calculated from the 6-month running mean of the monthly mean DMA8 ozone values at a given site. This metric was selected because it aligns with the ozone metric used by Turner et al. (2016) to quantify the impact of long-term ozone exposure on human mortality. Hereinafter this quantity is simply referred to as “the ozone metric".

2. *Atmospheric chemistry model simulations*: We use output from models from phase 1 of the Chemistry-Climate Model Initiative (CCMI), downloaded from the Centre for Environmental Data Analysis (CEDA) database (http://archive.ceda.ac.uk). We chose four models (CHASER, GEOSCCM, MOCAGE and MRI-ESM1r1) because they report hourly ozone output (Table 1). These particular simulations were part of CCMI’s REF-C2 experiment (Morgenstern et al., 2017) which follows the World Meteorological Organization (2011) A1 scenario for ozone depleting substances, and RCP 6.0 for tropospheric ozone precursors, and aerosol and aerosol precursor emissions (Morgenstern et al., 2010) for the period 1960-2100. Even though the most appropriate experiment would have been the REF-C1SD, in which the models are nudged to the reanalysis meteorology and thus best represent the past in the observations, we use output from the REF-C2 simulation in this study, as the last year of the REF-C1SD was 2010, and would therefore not cover the most recent period where observations are available. However, the NOAA Geophysical Fluid Dynamics Laboratory (GFDL) AM3 model continued the simulation over the entire study period and was therefore selected for this analysis. In addition, we obtained output from the GEOS-5 nature run with chemistry (G5NR-Chem), provided by the NASA Global Modeling and Assimilation Office (GMAO), which we included in our analysis because of the model’s very fine horizontal resolution (Hu et al., 2018), but the output was only available for July 2013 to June 2014.

The output from each individual model is shown in supplemental Fig S-1. Note that NASA G5NR-Chem has the finest resolution of these models; accordingly we aim to produce our final product on the same $0.125^\circ \times 0.125^\circ$ grid. However, even at this resolution the output is not street-resolving and thus will not capture urban scale variability in the regions with the highest population density.

In order to compare model output to observations, we need to register model output and observations to a common grid. This registration enables us to quantify the differences between the models and observations. Previous attempts have usually relied on a variant from a general statistical interpolation framework to combine incompatible spatial data (Gotway and Young, 2002;

Fuentes and Raftery, 2005; Gelfand and Sahu, 2010; Berrocal et al., 2012; Nguyen et al., 2012). Due to the highly irregular locations of ozone monitors around the globe, we use a kriging technique to build a statistical model, interpolate the ozone distribution based on the surrogate, and then project the global surface onto a common grid.

2.2 Fusion of observations and models

Following is a description of our method for fusing observations and output from multiple global atmospheric chemistry models to produce a surface ozone product with maximized accuracy. This method is known as Measurement and Multi-Model Fusion (version 1), or M^{3}Fusion (v1), and the code accompanies this manuscript in supplementary material. We consider a general framework of uncertainty quantification consisting of the following components (Kennedy and O'Hagan, 2001; Chang and Guillas, 2019):

$$\begin{aligned}\text{Observation} &= \text{Reality} + \text{Random Error};\\ \text{Reality} &= \text{Model} + \text{Structured Bias},\end{aligned}$$

Since this equation requires matching components (observations and model output) on a common grid, we use the interpolated observations to estimate an optimized weight for each model by a L^2 norm (details are given later), which means that we expect the multi-model combination to capture the general pattern of the surface ozone distribution in terms of their joint predictive capability; and the model bias is considered as a model correction term. The difference between observation error and model bias is that the former term is assumed to be a normal noise with zero mean and constant variance; and the latter term is considered as a systematic and structured discrepancy (Williamson et al., 2015), which will be revealed as a spatial cluster across a poorly simulated region.

Due to this study's human health focus we do not consider ozone above the data-sparse oceans. Above land, large observational gaps are present across Africa, the Middle East, South America, and South and Southeast Asia, where the spatial interpolation is generally too uncertain to yield a reliable surface ozone approximation. The ozone estimates in these regions must come from either models or distant observations, neither of which is ideal to solve this issue. As a compromise strategy we fill these gaps with a weighted model product evaluated by the interpolated ozone observations. We propose the following procedure to combine model output and observations for data integration:

1. *Interpolating irregularly located monitoring observations to the model output grid*: Kriging is a procedure used to statistically interpolate irregularly spaced and/or sparse observed data onto a regular and dense grid, based on a weighted average of the fitted surrogate model in the neighborhood of the grid. We assume that the global ozone distribution can be approximated by a Gaussian spatial process (GP) with a constant mean and Matérn covariance function (Stein, 2012). The GP fitting typically involves a cubic complexity, and thus is computationally expensive for large spatial data sets. Therefore several alternatives have been developed to address the large n issue by using a reduced set of data (Cressie and Johannesson, 2008; Banerjee et al., 2012; Liang et al., 2013), tapering the covariance between two grid points to zero if their distance is beyond a certain range (Furrer and Sain, 2009; Sang and Huang, 2012), and/or evaluating the

covariance only through the specification of a neighborhood system (also known as the Gaussian Markov random field) (Rue et al., 2009; Lindgren et al., 2011).

In this study we carry out the spatial interpolation by using the combination of the *integrated nested Laplacian approximations* (INLA) framework (Rue et al., 2009), and the *stochastic partial differential equation* (SPDE) technique (Lindgren et al., 2011), available as an R package (http://www.r-inla.org/) (Lindgren and Rue, 2015). The details of this technique are rather complex and the reader is referred to the original paper (Lindgren et al., 2011), however we describe the key component of this INLA-SPDE technique in the Appendix. INLA-SPDE spatial modeling has proven to be effective in a wide range of applications (Cameletti et al., 2013; Shaddick and Zidek, 2015; Heath et al., 2016; Liu and Guillas, 2017; Rue et al., 2017). We chose this technique because it manages a fairly large and complex spatial field in a relatively efficient way (Rue and Held, 2005), and allows an extension for nonstationarity on the sphere (Bolin and Lindgren, 2011; Chang et al., 2015). Notably, a recent study elaborately compared dozens of spatial modeling approaches, and the results suggest that almost all of these approaches can achieve a similar performance in terms of their predictive accuracy, albeit with very different computation times (Heaton et al., 2018). Therefore, we expect that the choice of spatial modeling approach is not the most crucial component in our data fusion process as long as the analysis is carried out in a rigorous way (i.e. through the statistical model selection and diagnostics). To differentiate this result from the actual observations in the TOAR database, we refer to this interpolated surface as the "spatially interpolated ozone".

We carry out the statistical interpolation via the following steps: (1) calculate the ozone metric at each TOAR site and for every year in 2008-2014; (2) perform the statistical interpolation using all available sites with their exact coordinates, and project the surface onto a $0.125° \times 0.125°$ spherical grid for every year; (3) average these surfaces over the 7 years to yield an observation-based present-day ozone distribution. We expect that this aggregation will smooth out at least some of the potential uncertainties. The kriging can be seen as a nonparametric regression problem, therefore a statistical assessment of fitted quality must be considered to select the best representation to the data (Hoeting et al., 2006). Further details on the statistical model selection procedure are provided in Appendix A.

We use a bilinear interpolation to smooth model output from coarser resolution to a $0.125° \times 0.125°$ grid (Jun et al., 2008). The ozone metric for each model was calculated for each single grid in each year, then averaged over 2008-2014 (except for NASA G5NR-Chem, which was already in fine resolution, but only available for 1 year).

2. *Weighting model output against spatially interpolated ozone by region*: The next step is to create an intermediate "multi-model composite". We divide the global land surface into 8 regions (see Fig 1), roughly matching the continental outlines or major population regions. We adopt this regional approach because global models vary in their ability to simulate ozone in different regions of the world. Next we regress the observations on multi-model output by a constrained least square approach within each of the eight regions. Let s_g be the grid cell at resolution $0.125° \times 0.125°$, $\hat{y}(s_g)$ be the interpolated observations, and $\{\eta_k(s_g); k = 1, \ldots, 6\}$ be the model output registered onto the same grid from the six

models considered in this paper (table 1). The optimization equation is based on a constrained least squares approach:

$$\underset{\{\alpha_r,\beta_{rk};k=1,\ldots,6\}}{\text{minimize}} \sum_{s_g \in \text{Region } r} \left(\hat{y}(s_g) - \alpha_r - \sum_{k=1}^{6} \beta_{rk} \eta_k(s_g) \right)^2, \tag{1}$$

subject to $\sum_{k=1}^{6} \beta_{rk} = 1$ and $\beta_{rk} \geq 0$.

where α_r is a constant that allows adjustment to the overall (regional) underestimation or overestimation, β_{rk} is an optimal weight for the k-th model in region r. Note that since the interpolated observations and models use the same ozone metric with the same units, we constrain the weights to be positive and sum to 1 for a better physical interpretability, such that the most accurate models receive the higher weight. The offset term α_r is aimed to adjust the overall residuals between the observation field and the multi-model composite into zero mean in each region (regardless of the spatial pattern), therefore if two spatial fields share a great similarity in terms of their spatial curvatures, but the overall means are different, this term can fill the gap of the overall mean difference between the two fields. The weights are optimized in terms of the squared distance between the interpolated ozone and multi-model output. A different criterion of optimization, such as mean absolute error, can be established accordingly.

Due to the sparsity of stations in many regions, we use a pre-defined geometric boundary to differentiate regions. A more meaningful physical boundary (i.e., regions with similar chemical regimes, or major features such as deserts, mountain ranges or water bodies) might be determined using a cluster analysis technique (Hyde et al., 2018), but such a step is beyond the scope of this paper.

Since we partition the global land surface into eight regions and evaluate the models individually, inevitably there will be disjointed boundaries between regions. The boundaries between North and South America, or between East Asia and Oceania, fall mostly in the oceans, so we do not need to adjust these regions. However, we should make an adjustment to disjointed boundaries that fall across inhabited areas (see supplementary Fig. S-2 for the illustration). As an example of our method, consider the boundary between East Asia and Russia near 50°N. We increase the northern boundary of East Asia to 55°N and decrease the southern boundary of Russia to 45°N, to create an overlapping intersection, and then fit cubic splines (performed for each grid cell) with knots placed at every 2 degree grid cell (Wood et al., 2008). This smoothing is carried out using a low rank Gaussian process by the default penalized least square from the function “gam” in the R package mgcv (Kammann and Wand, 2003; Wood, 2017), following the examples of Wood and Augustin (2002). The purpose is to merely avoid a sharp and unrealistic (geometric) transition between three regions and to efficiently smooth out the discontinuity, performed in a regular spaced grid only around the geometric boundary. Any region away from the geometric boundary will not be affected by this smoothing, which should be considered as a blending of multiple models without any attempt of bias correction. It should be noted that the INLA-SPDE technique in step 1 is applied to the observations, while the smoothing spline is only applied to the boundaries between regions of the model composite, not directly involving any observations.

We adopted a regression weighting approach that only accounts for the mean spatial fields of the interpolated ozone and model output, rather than the underlying associated uncertainty. We take this approach due to the prohibitive size of high resolution output (over 1 million output points for each model), but also due to the lack of a thorough investigation regarding the ideal method for combining models based on different sources of uncertainty. For example, the interpolation uncertainty can be quantified easily through the posterior distribution and considered to be related to measurement error (small scale) or sparse sampling across a region (large scale), however, model uncertainty is a different concept altogether that could result from input uncertainty (e.g. air pollution emissions inventories), or limitations of the transport and chemistry mechanisms within the model (Brynjarsdóttir and O'Hagan, 2014). The current interest of this study focuses on a better estimate of mean ozone exposure. Explicit quantification of different sources of model uncertainty and incorporation of this information into the data fusion process presents another level of complexity that cannot be tackled until model uncertainties are better characterized. Young et al. (2018) provide a current overview of chemistry-climate modelling and discuss the challenges of improving models in light of so many uncertainties.

3. *Correcting multi-model bias in areas close to observations*: A common practice of studying the model discrepancy in the spatial fields is to fit a statistical model for their differences from observations on the whole spatial domain, to see whether or not these residuals reveal any structured spatial pattern (Jun and Stein, 2004; Sang et al., 2011). If the model adequately simulates the ozone distribution (up to a level shift and a scale factor), then there is no relevant information in these residuals. On the other hand, if the model does not properly represent the local structure, then the residuals should exhibit a signal of the discrepancy in that region (Guillas et al., 2006; Williamson et al., 2015). However, in our case the regular grid observation field is obtained from spatial kriging, such that in many data sparse regions we don't actually have observed ozone, which prevents us from correcting the model in these regions. Instead, we conduct a limited model bias correction based on the distance to the nearest monitoring station, but we ignore the differences between the multi-model composite and the interpolated observations in the sparsely monitored regions. In our approach we only correct the output grid where there is at least one observational station within a 2 degree radial distance of the grid cell in question (i.e. the distance to the nearest station is less than 2°). We then end up using

$$\begin{cases} \hat{y}(s_g), \text{ if a grid cell } s_g \text{is within a 2 degree radial distance of the station;} \\ \alpha_r + \sum_{k=1}^{6} \beta_{rk}\eta_k(s_g), \text{otherwise,} \end{cases}$$

to generate our high resolution global surface ozone estimate. Given the limited availability of observations worldwide, we were only able to apply this final bias correction to 14.4% of the globe's land area. We refer to the final outcome as the "fused surface ozone product".

3 Results

3.1 Mapping and uncertainty

Ground based measurements were available from 4766 stations reported in the TOAR database (Schultz et al., 2017). To illustrate the spatial coverage of the database, Fig 1 shows the ozone metric discretized to a $2^\circ \times 2^\circ$ grid (a finer resolution will be too obscure for illustrative purposes), averaged over the period 2008-2014. This figure also shows our regionalized classification, including Africa, North America, South America, East Asia, Southeast and central Asia, Europe, Oceania, and Russia. Note that dense station networks are found in North America, Europe and East Asia (mostly in Japan and South Korea), while monitoring sites are more widely scattered across the remaining regions. The highest average ozone levels are found at sites in China, South Korea, Japan, Taiwan, India, Greece, California and Mexico City.

Fig 2(a) shows the spatially interpolated surface in each cell. For each grid cell, there is an underlying (posterior) probability distribution which incorporates information about the interpolation uncertainty. Fig. 2(b) shows the half-width of the 95% posterior credible interval in each cell (Shaddick et al., 2018). From the spatial pattern of uncertainty, we can see that relatively higher uncertainties are expected in Africa, the Middle East, South Asia and Russia, regions with very limited observations; lower uncertainty is associated with regions with a dense station network, such as North America and Europe. Due to the limitations of spatial kriging in a sparsely monitored region, the observations are often interpolated across very great distances, such as in South America, Africa and Central Asia. This method is not ideal, and instead, information from models can be used to fill in the blanks.

The ozone metric for each model was calculated for each individual grid cell in each year, then averaged over 2008-2014, and registered to the common $0.125^\circ \times 0.125^\circ$ grid (except for NASA G5NR-Chem, which was already in fine resolution, but only available for 1 year). Fig 3(a) shows the surface ozone metric which results from the simple ensemble average of the 6 models. It was generated from bilinear interpolation of the ozone metric on the standard output grid, by calculating the same metric for each grid cell in each year, averaging over 2008-2014, and then averaging over the 6 models. We refer to this product as the "multi-model mean", and we use it to validate our final product, which should outperform not only each individual model, but also the multi-model mean.

Averaging all 6 models captures the large scale variations of the ozone distribution, however, many regions in northern mid- and low latitudes are biased high compared to the observations in the TOAR database. A simple approach to address the uncertainty in the multi-model mean is to calculate the standard deviation for each grid cell from the different models, as shown in Fig 3(b). Higher model uncertainties across South Africa and the Middle East match the pattern of the interpolation uncertainty in Fig. 2(b), and lower model uncertainties occur in regions with dense station networks. These findings suggest that the multi-model mean uncertainty can also reflect the current limited understanding of surface ozone in regions with limited or no observations.

It should be noted that the spatially interpolated observations are smoother in regions with fewer sites, and reveal a more detailed structure in regions with a dense station network. In contrast the multi-model mean is more noisy. Even though we average across multiple years and multiple models, the resulting ozone metric can still be noisy because it is calculated at each grid cell independently. In order to make maximum use of the skill of each model, we restrict the model evaluation to the regional scale in the next section.

3.2 Regional model evaluation and Multi-model Composite

To evaluate the performance of each model in a given region, we calculate the mean differences over all grid cells within the region and summarize them with the root mean square error (RMSE). Let $\hat{y}(s_g)$ be the spatially interpolated observations, and $\{\eta_k(s_g); k = 1, \ldots, 6\}$ be the output corresponding to the six ensemble models considered in this paper, then the (normalized) RMSE is given by

$$\text{RMSE}_{rk} = \sqrt{\frac{\sum_{s_g \in Region\ r} \left(\eta_k(s_g) - \hat{y}(s_g)\right)^2}{n}},$$

where n is the number of grid cells in a given region r. The first part of Table 2 shows the RMSE statistics for each model by region. The reliability of such an evaluation is limited by the station density in a given region, with greater reliability in a dense network (e.g. USA) and less reliability in a sparse network (e.g. Africa, South America or Australia). On average, CHASER, GEOSCCM, and G5NR-Chem have the lowest biases in multiple regions; GFDL-AM3 and MRI-ESM1r1 also show low mean biases in certain regions, such as America and Europe. However, larger model biases can be found in Africa, East, and South Asia.

We next select three regions with extensive monitoring: North America, Europe and East Asia. Fig 4 shows the differences between the spatially interpolated observations and model output in North America. A consistent under-estimation can be found in the Mexico City region for all models. A clear over-estimation is also found across much of the eastern USA, as well as the western USA and Canada, except for CHASER which shows a mild under-estimation in these regions; In Europe (Fig 5), the models show mild levels of over-estimation across most of the region, especially for Italy . In East Asia (Fig 6), the models show a major bias across East China, and a similar bias pattern across the entire region, although the bias amplitude is smaller for GEOSCCM. However, since the observations are relatively sparse in mainland China, the large scale of these estimated biases might be an interpolation artifact.

We argue that the credibility of the model is not entirely decided by the RMSE (i.e. the mean difference): the smoother the difference plots, the easier it is to carry out the model bias correction. Indeed, the observations and model output are not expected to match point by point. We should also expect the model to capture the general pattern of the spatial distribution, rather than a point-wise agreement.

The estimated weights from the constrained least squares (Eq 1) are given in the second part of Table 2. Due to fixed underlying spatial structures, this approach tends to give greater weight to a single model (i.e. $\geq 50\%$), the one which provides the best match between its spatial structure and the observational field (e.g., G5NR-Chem in North America). Note that this approach disfavors noisy spatial structure, therefore the algorithm gives low weights to MOCAGE, for several reasons. First, the MOCAGE ozone field has not been smoothed by interpolation since it is already produced on the MOCAGE model grid, whereas all other models are interpolated. Secondly, MOCAGE uses a more complete tropospheric chemical scheme with a larger range of species (77 tropospheric species) and has generally a higher reactivity compared to most CCMs (Voulgarakis et al., 2013). Thus, it tends to provide more temporal and spatial variability. Note that our optimization algorithm estimates the weights according to the similarity of the spatial structures between the interpolated surface and each model. In regions with sparse monitoring the kriged surface can be greatly affected by a few scattered stations, therefore we cannot use the resulting weights to evaluate the actual model performance in these regions.

The last column of Table 2 shows the averaged and combined RMSEs from the equal weights and the constrained weights. A reduced overall bias can be generally achieved from the constrained weights. This approach suggests that even if a model has a large mean error (e.g. GFDL-AM3), it can still be a good simulation if it produces a spatial pattern and curvature similar to the observation field. A constant offset α_r in the optimization Eq (1) is included to remove the overall bias over each region, such that the residuals from the optimization have a zero mean. On the other hand, if we do not include α_r in the equation, GFDL-AM3 will have a smaller weight in the optimization, and CHASER, GEOSCCM and G5NR-Chem will dominate most of these regions (not shown).

We combine all models according to the optimum weights from each region for each model. Fig 7(a) shows a map of the multi-model composite, a weighted blend of the 6 models, with the weighting calculated separately for each continent. Models with greater simulation skill receive higher weighting. The result reveals a systematic adjustment to the large scale overestimation from the ensemble mean in Fig 3(a). This demonstration of a general high bias among the models argues against using the simple ensemble model mean for estimating surface ozone. However, when compared to the TOAR observations, the multi-model composite still has clear local biases.

3.3 Local bias correction

The last step of producing the final fused surface ozone product is to apply a bias correction to our multi-model composite, limited to just those areas in close proximity to ozone observations. Ideally we would like to apply a bias correction according to raw observations, but most stations are not exactly located on the model grid coordinates (even at $0.125° \times 0.125°$ resolution). Therefore, to carry out a statistical bias correction on a particular grid, we need to consider the number of nearby stations and the distance to each station. All these considerations aim to deduce a single correction value on a single grid, and thus we are still faced with implementing statistical interpolation. To avoid adding another level of complexity, we set the final fused product to be exactly equal to the spatially interpolated ozone field within 2 degrees of an observation, as the spatially interpolated ozone field has already accounted for all observations. Due to the global sparseness of observations, about 85% of model grid cells over land were not affected by this bias correction. After bias correcting the multi-model composite grid cells within 2-degrees of a TOAR observation site, an immediate benefit is seen for the USA, Mexico City, Italy and South Korea (see Fig 7(b)).

The choice of the correction range, in this case 2 degrees, is a ad hoc decision; we also present results with different correction ranges in supplementary Fig S-3 and S-4. When the radius of influence of the TOAR observations is increased to 5 or more degrees the greatest impact is seen for the Mexico City region and eastern China. An increase of correction range is not ideal because it extrapolates the Mexico City ozone values into the less populated regions of Mexico. Increasing the radius to 5 or more degrees does not improve upon the RMSE associated with 2 degrees. Therefore accepting the 2-degree bias correction over other ranges is subjective.

The fused product can be evaluated in terms of spatial correlation using the variogram which assumes that spatial correlation is not a function of absolute location, but only a function of distance (i.e., stationarity). Since spatial variability and continuity from the models are the result of geophysical processes represented by mathematical equations, the variogram must be customized for each field. In addition, the extremely large size of the model output prohibits us from carrying out a standard empirical variogram analysis, which requires calculating the variance of the difference between all pair-wise grid cells.

Nevertheless, we provide examples of omnidirectional variograms for the spatial field in North America from each model and product in supplementary Fig. S-5. The standard variogram analysis focuses on the following three parameters: (1) the nugget (variance at zero distance, which represents a sub-grid variation), which is similar for all cases; (2) the sill (total variance of a field), where the variogram value reaches a maximum and levels off; (3) the range (a distance where the sill is reached, and beyond that there is no longer spatial correlation). Note that a continuously increasing variogram indicates the evidence of non-stationarity in the field, which is the case for SPDE, an issue that we have accounted for. The variogram peak is about 35-40 degrees for the models. The result is very similar for G5NR-Chem, GEOSCCM and GFDL-AM3, while CHASER and MRI-ESM show a larger variance in the spatial field. The reason is that the latter two models produce low ozone in the high latitude region over Canada (see supplementary Fig. S-1), but the former three models simulate relatively higher ozone in the same region, and this difference is reflected by the total variance. Even though North America has one of the most extensive monitoring networks in the world, some of the remote areas (mostly in Canada) are mainly described by the model output in the final fused product. Therefore the variogram of the fused product is likely adjusted toward the remote areas of Canada as simulated by G5NR-Chem, which provided the largest weighting in North America).

3.4 Validation of the results

Since the raw observations are the only reliable source for validating our results, we align each model grid to observed locations for evaluating the predictive performance. The RMSE of the residuals from all observations in 2008-2014 are displayed in Table 3. Note that since the global network of monitoring stations is heavily weighted by North America, Europe, South Korea and Japan, these numbers are not representative of the sparsely monitored regions. We compare the fused surface ozone results to the simple multi-model mean from all 6 models. Our interim product, i.e. the multi-model composite, is also compared in Table 3.

Our multi-model composite outperforms the multi-model mean in terms of lowest mean predicted error. Based on the spatially interpolated observations, the resulting multi-model composite takes advantage of the strengths of each model, and achieves a better accuracy. This result proves that our approach is effective, since our interim product has already improved upon the simple multi-model mean. The bias correction further reduces the residuals: this is expected because the spatial kriging algorithm is designed to minimize the difference to observations, thus it has the lowest RMSE (this value is the same for the kriging result and the fused product since we apply the correction based on observed locations). The RMSE of approximately

5 ppb may represent the interannually varying meteorological influence during the years 2008-2014. If this is the case, then 5 ppb may approximate the minimal RMSE that can be achieved in a multi-year analysis.

In summary, the simple multi-model mean method may perform fairly well at the continental or regional scale, but does not provide an accurate representation of the sub-regional structure, this is of course a limitation on the use of coarse model resolutions. The weighting applied during the construction of the multi-model composite improved the accuracy but the effect could be limited, because many small-scale processes are not (yet) resolved by the models. To alleviate the discrepancy further, a statistical method based on local observations is applied to correct the bias. The advantage of our fused surface ozone product over the simple multi-model mean can be clearly seen in Figure 8. When interpreting the fused product the reader should consider the following: (1) For a region with an extensive monitoring network, such as the USA, a detailed bias correction can be achieved. We can utilize the observations to accurately reflect many local features (i.e., sub-grid variations) as shown in the ozone pollution hot-spots of southern California and Mexico City. However it should be noted that this improvement is due to local bias correction, instead of model weighting; (2) For regions with large observational gaps, such as South America, Africa or Russia, the spatial difference between the fused product and the multi-model mean is rather featureless, because the model weighting can only adjust the overall regional mean according to a few monitoring sites, and cannot address the local variations. Filling large data gaps with the intermediate multi-model composite can indeed avoid the influence of preferential sampling (Diggle et al., 2010; Shaddick and Zidek, 2014), but it is still subject to a high uncertainty due to lack of data.

4 Discussion and Conclusions

In this article we present a flexible framework to incorporate observations and multiple models for providing an improved estimate of the global surface ozone distribution. Combining multivariate spatial fields in the estimation of ozone distribution is an extension of both the conventional multi-model ensemble approach (i.e. simple average) and a statistical bias correction approach, and was found to improve the prediction of surface ozone. In summary our approach has the following properties:

1. The multi-year average enables us to reduce the meteorological influence on surface ozone. An extension of this method to time-resolved multi-annual fields can be expected to capture the interannual variability (Shaddick and Zidek, 2015), however such an endeavor would be highly computationally demanding in such a fine resolution setting.

2. The INLA-SPDE interpolation framework allows for modeling of potential nonstationarity in the spatial processes.

3. Regional model evaluation facilitates a feature selection for multiple competing atmospheric models.

4. Local bias correction of the multi-model composite only at a limited range of grid cells avoids using the spatially interpolated ozone field in regions associated with higher levels of uncertainty.

5. For the regions with dense monitoring networks (such as North American, Europe, South Korea and Japan), the final fused product was obtained mainly from the interpolation of observations; elsewhere the final product relied on the multi-model composite through an optimized weight from each model.

Human health studies typically adopt a fine grid resolution, such as a $0.1^{\circ} \times 0.1^{\circ}$ grid product, for matching to the gridded world population database. Even though the spatial kriging surrogate can produce the predicted value at any resolution, the accuracy of the fused surface ozone product is still limited by the density of observations around that point, and by the resolution of the global model output. Regarding future improvements two key developments can be expected to yield a better estimation of the global surface ozone distribution: Firstly, we can include more simulators for increased leverage. Another way to increase the estimation accuracy is to expand ozone monitoring networks across sparsely sampled regions (Sofen et al., 2016; Schultz et al., 2017; Weatherhead et al., 2017).

The application of our methodology focuses on, but is not limited to, a particular ozone metric relevant for quantifying the impact of long-term ozone exposure on human health. We expect that this framework could also be applied to other ozone metrics relevant to crop production or natural vegetation (Lefohn et al., 2018; Mills et al., 2018), or any other trace gas, provided adequate in situ observations are available for model evaluation.

In general, atmospheric chemistry model estimates of surface ozone levels are biased high, as demonstrated by a comparison of the annual mean surface ozone produced by the ACCMIP (Atmospheric Chemistry and Climate Model Intercomparison Project) multi-model ensemble to the TOAR Surface Ozone Database (see Figure 6 of Young et al. (2018)). This analysis has shown that the high bias is also prevalent among models when employing an ozone metric that focuses on the high end of the ozone distribution (Figure 8). Similarly, Shindell et al. (2018) compared the NASA GISS-E2 model to observed values of annual mean DMA8, and concluded that the model was biased high by 25%. Given the common tendency for models to over-estimate surface ozone, the methodology developed by this paper can be used to improve the accuracy of model output, either for individual models or for multi-model ensembles.

Code and data availability. The sources of the TOAR data and the output from 4 CCMI models are listed in Section 2.1; the output from the GFDL-AM3 model is archived at GFDL and is available to the public upon request to Meiyun Lin; G5NR-Chem model outputs are available for download at https://portal.nccs.nasa.gov/datashare/G5NR-Chem/Heracles/12.5km/DATA or can be accessed through the OpenDAP framework at the portal https://opendap.nccs.nasa.gov/dods/OSSE/G5NR-Chem/Heracles/12.5km; All computations in our methodology are implemented in R (R Core Team, 2013). The relevant code can be found in R packages for statistical interpolation (R-INLA, Lindgren and Rue (2015)), quadratic programming (limSolve) and spline smoothing (mgcv, Wood (2017)). The R code accompanies this manuscript on its associated GMD webpage.

Appendix A: Spatial modeling using the INLA-SPDE approach

In this paper the aim of spatial interpolation is to use (discretized) monitoring observations to build a statistical surrogate model for estimating the ozone distribution over the whole domain on a sphere. We assume that this ozone distribution follows a Gaussian process (GP). A GP is a collection of random variables such that any subset of the observations has a joint Gaussian distribution. It has been widely used in many applications as a machine learning algorithm (Rasmussen and Williams, 2006). In this section we briefly introduce the GP model with a focus on spatial kriging. The GP is a popular choice in spatial statistics because it allows modeling of fairly complicated functional forms, and it also provides a prediction and associated uncertainty at any new location. A common limitation of this interpolation is that the resulting distribution of estimated uncertainty will be lower around individual stations or within dense monitoring networks, and higher in sparsely monitored regions.

Let Y denote an n-vector of ozone observations measured at monitoring sites $\mathbf{s}$, then a statistical model for the spatial field can be expressed as: $Y = f(\mathbf{s}) + \varepsilon$, i.e. the model comprises a smooth GP spatial process $f(\mathbf{s})$, capturing spatial association, and an independent normal error ε, which follows a normal error $N(0,\sigma^2)$. This error term can accommodate potential measurement error; on the other hand, kriging without measurement error is usually used for the surrogate of a deterministic model (i.e. the same input always produces the same output), also known as an emulator (e.g. Conti and O'Hagan (2010)).

The specification of a GP is through its mean function and covariance function, denoting by $f(\mathbf{s}) \sim GP(m(\mathbf{s}), c(\mathbf{s},\mathbf{s}'))$. To reduce computational intensity, the mean function can be assumed to be a constant $m(\mathbf{s}) = \mu$, thus the resulting spatial distribution is completely defined by the covariance function. A covariance function characterizes correlations between different locations in the spatial process, it is the crucial component in a GP, as it represents our assumptions about the latent field from which we wish to build a surrogate. Specifically, we use the Matérn covariance function, which is a flexible covariance structure and widely used in spatial statistics (Hoeting et al., 2006; Jun and Stein, 2007, 2008). With the shape parameter $\nu > 0$, the scale parameter $\kappa > 0$, and the marginal precision $\tau^2 > 0$, the covariance structure can be written as:

$$c(\mathbf{h}) = \frac{2^{1-\nu}}{4\pi\kappa^{2\nu}\tau^2\Gamma(\nu+1)}(\kappa\|\mathbf{h}\|)^\nu K_\nu(\kappa\|\mathbf{h}\|), \mathbf{h} \in \mathbb{S}^2,$$

where $\mathbf{h}$ denotes the distance between any two locations: $\mathbf{h} = \mathbf{s} - \mathbf{s}'$, Γ is a gamma function, and K_ν is the modified Bessel function of the second kind of order $\nu > 0$. The scale parameter κ controls the rate of decay of the correlation between two locations as distance increases. Smaller values of κ, allow for longer ranges over which two sites can be correlated. The smoothness parameter ν can be seen as the determining behavior of the autocorrelation for observations that are separated by a small distance.

The major disadvantage of using a GP is the computational complexity, which typically involves a cubic complexity in the number of data points, usually denoted as $O(n^3)$. Several attempts have been made to reduce the computational burden: e.g. Cressie and Johannesson (2008), Rue et al. (2009), Banerjee et al. (2012) and Gramacy and Apley (2015). Lindgren et al. (2011) introduced a popular approach in which the Matérn covariance can be approximated by the solution of certain stochastic partial differential equations (SPDE). According to Lindgren et al. (2011), a GP process $f(\mathbf{s})$ with Matérn covariance on a sphere is the solution of the following stationary SPDE:

$$(\kappa^2 - \Delta)^{(\nu+1)/2}\tau f(\mathbf{s}) = \mathcal{W}(\mathbf{s}),$$

where Δ is the Laplace operator and $\mathcal{W}$ is the Gaussian white noise. The core implication of this mathematical relationship is that an efficient algorithm for solving this SPDE can be applied to approximate the GP (Lindgren et al., 2011).

This INLA-SPDE technique also enables us to quantify the level of nonstationarity in a spatial field by employing basis function representations for both κ and τ (i.e. these quantities are constants in a stationary field). To obtain basic identifiability,

$\kappa(\mathbf{s})$ and $\tau(\mathbf{s})$ are taken to be positive, and their logarithm can be represented as:

$$\log\kappa(\mathbf{s}) = \sum_{k=1}^{p} \theta_k^{\kappa} \psi_k(\mathbf{s}) \quad \text{and} \quad \log\tau(\mathbf{s}) = \sum_{k=1}^{p} \theta_k^{\tau} \psi_k(\mathbf{s}), \tag{A1}$$

where $\{\psi_k(\mathbf{s})\}$ is a set of spherical harmonics. The coefficients $\{\theta_k^{\kappa}\}$ and $\{\theta_k^{\tau}\}$ represent local variances and correlation ranges (Bolin and Lindgren, 2011; Lindgren et al., 2011). A larger number of basis functions permits the representation of smaller local features.

We now illustrate a series of statistical model fits to select the best predictive ability of the SPDE model. To choose the maximum number of basis functions for the parameters κ and τ in equation A1, model selection techniques must be used. We perform the model selection based on the following criteria:

- RMSE (root-mean-square error): measure of the overall mean difference between predicted values and the observed values;
- DIC (deviance information criterion): the DIC is a measure to compare performance of statistical models by using a criterion based on a trade-off between the goodness of fit and the corresponding complexity of the model. Smaller values of the DIC indicate a better balance between complexity and a good fit;
- GCV (generalized cross validation): the mean residuals in a leave-one-out test. The model that minimizes the average predicted residuals over all the data is selected as the best model (Schneider, 2001).

We estimate 9 statistical models with different numbers of basis functions, presented in Table A1. The simplest model is a stationary Matérn model (we use basis number 0 to represent the κ and τ as constants). The best fit of all criteria occurs when the orders of the basis functions are increased from four to five. We therefore conclude that a model with five spatially varying basis functions is most appropriate for the TOAR observations.

Competing interests. The authors have no competing interests to declare.

Acknowledgements. This work was funded by the NASA Health and Air Quality Applied Sciences Team (grant #NNX16AQ80G).

References

Adachi, Y., Yukimoto, S., Deushi, M., Obata, A., andTaichu. Y. Tanaka, H. N., Hosaka, M., Sakami, T., Yoshimura, H., Hirabara, M., Shindo, E., Tsujino, H., Mizuta, R., Yabu, S., Koshiro, T., Ose, T., and Kitoh, A.: Basic performance of a new earth system model of the Meteorological Research Institute (MRI-ESM1), Pap. Meteorol. Geophys, 64, 1–18, https://doi.org/10.2467/mripapers.64.1, 2013.

Anenberg, S. C., Horowitz, L. W., Tong, D. Q., and West, J. J.: An estimate of the global burden of anthropogenic ozone and fine particulate matter on premature human mortality using atmospheric modeling, Environmental Health Perspectives, 118, 1189, 2010.

Banerjee, A., Dunson, D. B., and Tokdar, S. T.: Efficient Gaussian process regression for large datasets, Biometrika, 100, 75–89, https://doi.org/10.1093/biomet/ass068, 2012.

Berrocal, V. J., Gelfand, A. E., and Holland, D. M.: Space-time data fusion under error in computer model output: An application to modeling air quality, Biometrics, 68, 837–848, https://doi.org/10.1111/j.1541-0420.2011.01725.x, 2012.

Bolin, D. and Lindgren, F.: Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping, Ann. Appl. Stat., 5, 523–550, https://doi.org/10.1214/10-AOAS383, 2011.

Brauer, M., Amann, M., Burnett, R. T., Cohen, A., Dentener, F., Ezzati, M., Henderson, S. B., Krzyzanowski, M., Martin, R. V., Dingenen, R. V., van Donkelaar, A., and Thurston, G. D.: Exposure assessment for estimation of the global burden of disease attributable to outdoor air pollution, Environ. Sci. Technol., 46, 652–660, https://doi.org/10.1021/es2025752, 2012.

Brauer, M., Freedman, G., Frostad, J., van Donkelaar, A., Martin, R. V., Dentener, F., van Dingenen, R., Estep, K., Amini, H., Apte, J. S., Balakrishnan, K., Barregardh, L., Broday, D., Feigin, V., Ghosh, S., Hopke, P. K., Knibbs, L. D., Kokubo, Y., Liu, Y., Ma, S., Morawska, L., Sangrador, J. L. T., Shaddick, G., Anderson, H. R., Vos, T., Forouzanfar, M. H., Burnett, R. T., and Cohen, A.: Ambient air pollution exposure estimation for the global burden of disease 2013, Environ. Sci. Technol., 50, 79–88, https://doi.org/10.1021/acs.est.5b03709, 2015.

Braverman, A., Chatterjee, S., Heyman, M., and Cressie, N.: Probabilistic evaluation of competing climate models, Adv. Stat. Clim. Meteorol. Oceanogr., 3, 93–105, https://doi.org/10.5194/ascmo-3-93-2017, 2017.

Brynjarsdóttir, J. and O'Hagan, A.: Learning about physical parameters: The importance of model discrepancy, Inverse Problems, 30, 114 007, 2014.

Buser, C. M., Künsch, H. R., Lüthi, D., Wild, M., and Schär, C.: Bayesian multi-model projection of climate: bias assumptions and interannual variability, Clim. Dyn., 33, 849–868, https://doi.org/10.1007/s00382-009-0588-6, 2009.

Cameletti, M., Lindgren, F., Simpson, D., and Rue, H.: Spatio-temporal modeling of particulate matter concentration through the SPDE approach, AStA Adv. Stat. Anal., 97, 109–131, https://doi.org/10.1007/s10182-012-0196-3, 2013.

Cariolle, D. and Teyssèdre, H.: A revised linear ozone photochemistry parameterization for use in transport and general circulation models: multi-annual simulations, Atmos. Chem. Phys., 7, 2183–2196, https://doi.org/10.5194/acp-7-2183-2007, 2007.

Chandler, R. E.: Exploiting strength, discounting weakness: combining information from multiple climate simulators, Phil. Trans. R. Soc. A, 371, 20120 388, https://doi.org/10.1098/rsta.2012.0388, 2013.

Chang, K.-L. and Guillas, S.: Computer model calibration with large non-stationary spatial outputs: application to the calibration of a climate model, J. Roy. Stat. Soc. C, 68, 51–78, https://doi.org/10.1111/rssc.12309, 2019.

Chang, K.-L., Guillas, S., and Fioletov, V. E.: Spatial mapping of ground-based observations of total ozone, Atmos. Meas. Tech., 8, 4487–4505, https://doi.org/10.5194/amt-8-4487-2015, 2015.

Chang, K.-L., Petropavlovskikh, I., Cooper, O. R., Schultz, M. G., and Wang, T.: Regional trend analysis of surface ozone observations from monitoring networks in eastern North America, Europe and East Asia, Elem. Sci. Anth., 5, https://doi.org/10.1525/elementa.243, 2017.

Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., Freedman, G., Hubbell, B., Jobling, A., Kan, H., Knibbs, L., Liu, Y., Martin, R., Morawska, L., III, C. A. P., Shin, H., Straif, K., Shaddick, G., Thomas, M., van Dingenen, R., van Donkelaar, A., Vos, T., Murray, C. J. L., and Forouzanfar, M. H.: Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015, The Lancet, 389, 1907–1918, https://doi.org/10.1016/S0140-6736(17)30505-6, 2017.

Conti, S. and O'Hagan, A.: Bayesian emulation of complex multi-output and dynamic computer models, J. Stat. Plan. Inference, 140, 640–651, https://doi.org/10.1016/j.jspi.2009.08.006, 2010.

Cooper, O. R., Parrish, D. D., Ziemke, J. R., Balashov, N. V., Cupeiro, M., Galbally, I. E., Gilge, S., Horowitz, L., Jensen, N. R., Lamarque, J.-F., Naik, V., Oltmans, S. J., Schwab, J., Shindell, D. T., Thompson, A. M., Thouret, V., Wang, Y., and Zbinden, R. M.: Global distribution and trends of tropospheric ozone: An observation-based review, Elem. Sci. Anth., 2, https://doi.org/10.12952/journal.elementa.000029, 2014.

Cressie, N. and Johannesson, G.: Fixed rank kriging for very large spatial data sets, J. Roy. Stat. Soc. B, 70, 209–226, https://doi.org/10.1111/j.1467-9868.2007.00633.x, 2008.

Diggle, P. J., Menezes, R., and Su, T.-l.: Geostatistical inference under preferential sampling, J. Roy. Stat. Soc. C, 59, 191–232, 2010.

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., and Feng, Z.: Tropospheric Ozone Assessment Report: Present-day ozone distribution and trends relevant to human health, Elem. Sci. Anth., 6, https://doi.org/10.1525/elementa.291, 2018.

Fuentes, M. and Raftery, A. E.: Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models, Biometrics, 61, 36–45, https://doi.org/10.1111/j.0006-341X.2005.030821.x, 2005.

Furrer, R. and Sain, S. R.: Spatial model fitting for large datasets with applications to climate and microarray problems, Stat. Comput., 19, 113–128, https://doi.org/10.1007/s11222-008-9075-x, 2009.

Gaudel, A., Cooper, O. R., Ancellet, G., Barret, B., Boynard, A., Burrows, J. P., Clerbaux, C., Coheur, P. F., Cuesta, J., Cuevas, E., Doniki, S., Dufour, G., Ebojie, F., Foret, G., Garcia, O., Muños, M. J. G., Hannigan, J. W., Hase, F., Huang, G., Hassler, B., Hurtmans, D., Jaffe, D., Jones, N., Kalabokas, P., Kerridge, B., Kulawik, S. S., Latter, B., Leblanc, T., Flochmoën, E. L., Lin, W., Liu, J., Liu, X., Mahieu, E., McClure-Begley, A., Neu, J. L., Osman, M., Palm, M., Petetin, H., Petropavlovskikh, I., Querel, R., Rahpoe, N., Rozanov, A., Schultz, M. G., Schwab, J., Siddans, R., Smale, D., Steinbacher, M., Tanimoto, H., Tarasick, D. W., Thouret, V., Thompson, A. M., Trickl, T., Weatherhead, E. C., Wespes, C., Worden, H. M., Vigouroux, C., Xu, X., Zeng, G., and Ziemke, J. R.: 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, https://doi.org/10.1525/elementa.243, 2018.

GBD: Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013, The Lancet, 386, 2287–2323, https://doi.org/10.1016/S0140-6736(15)00128-2, 2015.

Gelfand, A. E. and Sahu, S. K.: Combining monitoring data and computer model output in assessing environmental exposure, in: Handbook of Applied Bayesian Analysis. Oxford University Press: Oxford, UK, pp. 482–510, 2010.

Gotway, C. A. and Young, L. J.: Combining incompatible spatial data, J. Am. Stat. Assoc., 97, 632–648, https://doi.org/10.1198/016214502760047140, 2002.

Gramacy, R. B. and Apley, D. W.: Local Gaussian process approximation for large computer experiments, J. Comput. Graph. Stat., 24, 561–578, https://doi.org/10.1080/10618600.2014.914442, 2015.

Guillas, S., Tiao, G. C., Wuebbles, D. J., and Zubrow, A.: Statistical diagnostic and correction of a chemistry-transport model for the prediction of total column ozone, Atmos. Chem. Phys., 6, 525–537, https://doi.org/10.5194/acp-6-525-2006, 2006.

He, Y. and Xiu, D.: Numerical strategy for model correction using physical constraints, J. Comput. Phys., 313, 617–634, https://doi.org/10.1016/j.jcp.2016.02.054, 2016.

Heath, A., Manolopoulou, I., and Baio, G.: Estimating the expected value of partial perfect information in health economic evaluations using integrated nested Laplace approximation, Stat. Med., 35, 4264–4280, https://doi.org/10.1002/sim.6983, 2016.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D. W., Sun, F., and Zammit-Mangion, A.: A case study competition among methods for analyzing large spatial data, J. Agric. Biol. Environ. Stat., pp. 1–28, https://doi.org/10.1007/s13253-018-00348-w, 2018.

Hoeting, J. A., Davis, R. A., Merton, A. A., and Thompson, S. E.: Model selection for geostatistical models, Ecol. Appl., 16, 87–98, https://doi.org/10.1890/04-0576, 2006.

Hu, L., Keller, C. A., Long, M. S., Sherwen, T., Auer, B., Da Silva, A., Nielsen, J. E., Pawson, S., Thompson, M. A., Trayanov, A. L., Travis, K. R., Grange, S. K., Evans, M. J., and Jacob, D. J.: Global simulation of tropospheric chemistry at 12.5 km resolution: performance and evaluation of the GEOS-Chem chemical module (v10-1) within the NASA GEOS Earth system model (GEOS-5 ESM), Geosci. Model Dev., 11, 4603–4620, https://doi.org/10.5194/gmd-11-4603-2018, 2018.

Hyde, R., Hossaini, R., and Leeson, A. A.: Cluster-based analysis of multi-model climate ensembles, Geosci. Model Dev., 11, 2033–2048, https://doi.org/10.5194/gmd-11-2033-2018, 2018.

Jerrett, M., Burnett, R. T., Pope III, C. A., Ito, K., Thurston, G., Krewski, D., Shi, Y., Calle, E., and Thun, M.: Long-term ozone exposure and mortality, N. Engl. J. Med., 360, 1085–1095, https://doi.org/10.1164/rccm.201508-1633OC, 2009.

Josse, B., Simon, P., and Peuch, V.-H.: Radon global simulations with the multiscale chemistry and transport model MOCAGE, Tellus B, 56, 339–356, https://doi.org/10.1111/j.1600-0889.2004.00112.x, 2004.

Jun, M. and Stein, M. L.: Statistical comparison of observed and CMAQ modeled daily sulfate levels, Atmos. Environ., 38, 4427–4436, https://doi.org/10.1016/j.atmosenv.2004.05.019, 2004.

Jun, M. and Stein, M. L.: An approach to producing space–time covariance functions on spheres, Technometrics, 49, 468–479, https://doi.org/10.1198/004017007000000155, 2007.

Jun, M. and Stein, M. L.: Nonstationary covariance models for global data, Ann. Appl. Stat., 2, 1271–1289, https://doi.org/10.1214/08-AOAS183, 2008.

Jun, M., Knutti, R., and Nychka, D. W.: Spatial analysis to quantify numerical model bias and dependence: how many climate models are there?, J. Am. Stat. Assoc., 103, 934–947, https://doi.org/10.1198/016214507000001265, 2008.

Kammann, E. and Wand, M. P.: Geoadditive models, J. Roy. Stat. Soc. C, 52, 1–18, https://doi.org/10.1111/1467-9876.00385, 2003.

Kennedy, M. C. and O'Hagan, A.: Bayesian calibration of computer models, J. Roy. Stat. Soc. B, 63, 425–464, https://doi.org/10.1111/1467-9868.00294, 2001.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A.: Challenges in combining projections from multiple climate models, J. Clim., 23, 2739–2758, https://doi.org/10.1175/2009JCLI3361.1, 2010.

Lefohn, A. S., Malley, C. S., Smith, L., Wells, B., Hazucha, M., Simon, H., Naik, V., Mills, G., Schultz, M. G., Paoletti, E., De Marco, A., Xu, X., Zhang, L., Wang, T., Neufeld, H. S., Musselman, R. C., Tarasick, D., Brauer, M., Feng, Z., Tang, H., Kobayashi, K., Sicard, P., Solberg, S., and Gerosa, G.: Tropospheric ozone assessment report: Global ozone metrics for climate change, human health, and crop/ecosystem research, Elem. Sci. Anth., 6, https://doi.org/10.1525/elementa.279, 2018.

Liang, F., Cheng, Y., Song, Q., Park, J., and Yang, P.: A resampling-based stochastic approximation method for analysis of large geostatistical data, Journal of the American Statistical Association, 108, 325–339, 2013.

Lin, M., Fiore, A. M., Horowitz, L. W., Cooper, O. R., Naik, V., Holloway, J., Johnson, B. J., Middlebrook, A. M., Oltmans, S. J., Pollack, I. B., Ryerson, T. B., Warner, J. X., Wiedinmyer, C., Wilson, J., and Wyman, B.: Transport of Asian ozone pollution into surface air over the western United States in spring, J. Geophys. Res. Atmos., 117, https://doi.org/10.1029/2011JD016961, 2012.

Lin, M., Horowitz, L. W., Oltmans, S. J., Fiore, A. M., and Fan, S.: Tropospheric ozone trends at Mauna Loa Observatory tied to decadal climate variability, Nat. Geosci., 7, 136–143, https://doi.org/10.1038/NGEO2066, 2014.

Lin, M., Horowitz, L. W., Payton, R., Fiore, A. M., and Tonnesen, G.: US surface ozone trends and extremes from 1980 to 2014: quantifying the roles of rising Asian emissions, domestic controls, wildfires, and climate, Atmos. Chem. Phys., 17, 2943–2970, 2017.

Lindgren, F. and Rue, H.: Bayesian spatial and spatiotemporal modelling with R-INLA, J. Stat. Softw., 63, https://doi.org/10.18637/jss.v063.i19, 2015.

Lindgren, F., Rue, H., and Lindström, J.: An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach, J. Royal Stat. Soc. B, 73, 423–498, https://doi.org/10.1111/j.1467-9868.2011.00777.x, 2011.

Liu, X. and Guillas, S.: Dimension reduction for Gaussian process emulation: an application to the influence of bathymetry on tsunami heights, SIAM/ASA J. Uncertain. Quantif., 5, 787–812, https://doi.org/10.1137/16M1090648, 2017.

Malley, C. S., Henze, D. K., Kuylenstierna, J. C., Vallack, H. W., Davila, Y., Anenberg, S. C., Turner, M. C., and Ashmore, M. R.: Updated global estimates of respiratory mortality in adults ≥ 30 years of age attributable to long-term ozone exposure, Environmental Health Perspectives, 125, 2017.

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., and Xu, X.: Tropospheric Ozone Assessment Report: Present-day tropospheric ozone distribution and trends relevant to vegetation, Elem. Sci. Anth., 6, https://doi.org/10.1525/elementa.302, 2018.

Morgenstern, O., Giorgetta, M. A., Shibata, K., Eyring, V., Waugh, D. W., Shepherd, T. G., Akiyoshi, H., Austin, J., Baumgaertner, A. J. G., Bekki, S., Braesicke, P., Brühl, C., Chipperfield, M., Cugnet, D., Dameris, M., Dhomse, S., Frith, S. M., Garny, H., Gettelman, A., Hardiman, S. C., Hegglin, M. I., Jöckel, P., Kinnison, D. E., Lamarque, J.-F., Mancini, E., Manzini, E., Marchand, M., Michou, M., Nakamura, T., Nielsen, J. E., Olivié, D., Pitari, G., Plummer, D. A., Rozanov, E., Scinocca, J. F., Smale, D., Teyssèdre, H., Toohey, M., Tian, W., and Yamashita, Y.: Review of the formulation of present-generation stratospheric chemistry-climate models and associated external forcings, J. Geophys. Res. Atmos., 115, https://doi.org/10.1029/2009JD013728, 2010.

Morgenstern, O., Hegglin, M. I., Rozanov, E., O'Connor, F. M., Abraham, N. L., Akiyoshi, H., Archibald, A. T., Bekki, S., Butchart, N., Chipperfield, M. P., Deushi, M., Dhomse, S. S., Garcia, R. R., Hardiman, S. C., Horowitz, L. W., Jöckel, P., Josse, B., Kinnison, D., Lin, M., Mancini, E., Manyin, M. E., Marchand, M., Marécal, V., Michou, M., Oman, L. D., Pitari, G., Plummer, D. A., Revell, L. E., Saint-Martin, D., Schofield, R., Stenke, A., Stone, K., Sudo, K., Tanaka, T. Y., Tilmes, S., Yamashita, Y., Yoshida, K., and Zeng, G.: Review of the global models used within phase 1 of the Chemistry–Climate Model Initiative (CCMI), Geosci. Model Dev., 10, 639–671, https://doi.org/10.5194/gmd-10-639-2017, 2017.

Nguyen, H., Cressie, N., and Braverman, A.: Spatial statistical data fusion for remote sensing applications, J. Am. Stat. Assoc., 107, 1004–1018, https://doi.org/10.1080/01621459.2012.694717, 2012.

Oman, L. D., Ziemke, J. R., Douglass, A. R., Waugh, D. W., Lang, C., Rodriguez, J. M., and Nielsen, J. E.: The response of tropical tropospheric ozone to ENSO, Geophys. Res. Lett., 38, https://doi.org/10.1029/2011GL047865, 2011.

R Core Team: R: A language and environment for statistical computing, 2013.

Rasmussen, C. E. and Williams, C. K. I.: Gaussian processes for machine learning, The MIT Press, Cambridge, MA, USA, 2006.

Rue, H. and Held, L.: Gaussian Markov random fields: theory and applications, CRC press, 2005.

Rue, H., Martino, S., and Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, J. Roy. Stat. Soc. B, 71, 319–392, https://doi.org/10.1111/j.1467-9868.2008.00700.x, 2009.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K.: Bayesian computing with INLA: a review, Annu. Rev. Stat. Appl., 4, 395–421, https://doi.org/10.1146/annurev-statistics-060116-054045, 2017.

Sang, H. and Huang, J. Z.: A full scale approximation of covariance functions for large spatial data sets, J. Roy. Stat. Soc. B, 74, 111–132, https://doi.org/10.1111/j.1467-9868.2011.01007.x, 2012.

Sang, H., Jun, M., and Huang, J. Z.: Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors, Ann. Appl. Stat., pp. 2519–2548, https://doi.org/10.1214/11-AOAS478, 2011.

Schneider, T.: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values, Journal of climate, 14, 853–871, 2001.

Schultz, M. G., Schröder, S., Lyapina, O., Cooper, O. R., Galbally, I., Petropavlovskikh, I., von Schneidemesser, E., Tanimoto, H., Elshorbany, Y., Naja, M., Seguel, R., Dauert, U., Eckhardt, P., Feigenspahn, S., Fiebig, M., Hjellbrekke, A.-G., Hong, Y.-D., Kjeld, P. C., Koide, H., Lear, G., Tarasick, D., Ueno, M., Wallasch, M., Baumgardner, D., Chuang, M.-T., Gillett, R., Lee, M., Molloy, S., Moolla, R., Wang, T., Sharps, K., Adame, J. A., Ancellet, G., Apadula, F., Artaxo, P., Barlasina, M., Bogucka, M., Bonasoni, P., Chang, L., Colomb, A., Cuevas, E., Cupeiro, M., Degorska, A., Ding, A., Fröhlich, M., Frolova, M., Gadhavi, H., Gheusi, F., Gilge, S., Gonzalez, M. Y., Gros, V., Hamad, S. H., Helmig, D., Henriques, D., Hermansen, O., Holla, R., Huber, J., Im, U., Jaffe, D. A., Komala, N., Kubistin, D., Lam, K.-S., Laurila, T., Lee, H., Levy, I., Mazzoleni, C., Mazzoleni, L., McClure-Begley, A., Mohamad, M., Murovic, M., Navarro-Comas, M., Nicodim, F., Parrish, D., Read, K. A., Reid, N., Ries, L., Saxena, P., Schwab, J. J., Scorgie, Y., Senik, I., Simmonds, P., Sinha, V., Skorokhod, A., Spain, G., Spangl, W., Spoor, R., Springston, S. R., Steer, K., Steinbacher, M., Suharguniyawan, E., Torre, P., Trickl, T., Weili, L., Weller, R., Xu, X., Xue, L., and Zhiqiang, M.: Tropospheric Ozone Assessment Report: Database and metrics data of global surface ozone observations, Elem. Sci. Anth., 5, https://doi.org/10.1525/elementa.244, 2017.

Seltzer, K. M., Shindell, D. T., and Malley, C. S.: Measurement-based assessment of health burdens from long-term ozone exposure in the United States, Europe, and China, Environmental Research Letters, 13, 104 018, 2018.

Shaddick, G. and Zidek, J. V.: A case study in preferential sampling: Long term monitoring of air pollution in the UK, Spatial Statistics, 9, 51–65, 2014.

Shaddick, G. and Zidek, J. V.: Spatio-temporal methods in environmental epidemiology, CRC Press, 2015.

Shaddick, G., Thomas, M. L., Green, A., Brauer, M., Donkelaar, A., Burnett, R., Chang, H. H., Cohen, A., Dingenen, R. V., Dora, C., Gumy, S., Liu, Y., Martin, R., Waller, L. A., West, J. J., Zidek, J. V., and Prüss-Ustün, A.: Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution, J. Roy. Stat. Soc. C, 67, 231–253, https://doi.org/10.1111/rssc.12227, 2018.

Shindell, D., Faluvegi, G., Seltzer, K., and Shindell, C.: Quantified, localized health benefits of accelerated carbon dioxide emissions reductions, Nature Climate Change, 8, 291, 2018.

Sofen, E. D., Bowdalo, D., and Evans, M. J.: How to most effectively expand the global surface ozone observing network, Atmos. Chem. Phys., 16, 1445–1457, https://doi.org/10.5194/acp-16-1445-2016, 2016.

Stainforth, D. A., Allen, M. R., Tredger, E. R., and Smith, L. A.: Confidence, uncertainty and decision-support relevance in climate predictions, Phil. Trans. R. Soc. A, 365, 2145–2161, https://doi.org/10.1098/rsta.2007.2074, 2007.

Stein, M. L.: Interpolation of spatial data: some theory for kriging, Springer Science and Business Media, 2012.

Stevenson, D. S., Dentener, F. J., Schultz, M. G., Ellingsen, K., Noije, T. P. C. V., Wild, O., Zeng, G., Amann, M., Atherton, C. S., Bell, N., Bergmann, D. J., Bey, I., Butler, T., Cofala, J., Collins, W. J., Derwent, R. G., Doherty, R. M., Drevet, J., Eskes, H. J., Fiore, A. M., Gauss, M., Hauglustaine, D. A., Horowitz, L. W., Isaksen, I. S. A., Krol, M. C., Lamarque, J.-F., Lawrence, M. G., Montanaro, V., Müller, J.-F., Pitari, G., Prather, M. J., Pyle, J. A., Rast, S., Rodriguez, J. M., Sanderson, M. G., Savage, N. H., Shindell, D. T., Strahan, S. E., Sudo, K., and Szopa, S.: Multimodel ensemble simulations of present-day and near-future tropospheric ozone, J. Geophys. Res. Atmos., 111, https://doi.org/10.1029/2005JD006338, 2006.

Strode, S. A., Rodriguez, J. M., Logan, J. A., Cooper, O. R., Witte, J. C., Lamsal, L. N., Damon, M., Van Aartsen, B., Steenrod, S. D., and Strahan, S. E.: Trends and variability in surface ozone over the United States, J. Geophys. Res. Atmos., 120, 9020–9042, https://doi.org/10.1002/2014JD022784, 2015.

Sudo, K., Takahashi, M., and Akimoto, H.: CHASER: A global chemical model of the troposphere 2. Model results and evaluation, J. Geophys. Res. Atmos., 107, https://doi.org/10.1029/2001JD001114, 2002a.

Sudo, K., Takahashi, M., Kurokawa, J., and Akimoto, H.: CHASER: A global chemical model of the troposphere 1. Model description, J. Geophys. Res. Atmos., 107, https://doi.org/10.1029/2001JD001113, 2002b.

Teyssèdre, H., Michou, M., Clark, H. L., Josse, B., Karcher, F., Olivié, D., Peuch, V.-H., Saint-Martin, D., Cariolle, D., Attié, J.-L., Nédélec, P., Ricaud, P., Thouret, V., van der A, R. J., Volz-Thomas, A., and Chéroux, F.: A new tropospheric and stratospheric Chemistry and Transport Model MOCAGE-Climat for multi-year studies: evaluation of the present-day climatology and sensitivity to surface processes, Atmos. Chem. Phys., 7, 5860, https://doi.org/10.5194/acp-7-5815-2007, 2007.

Turner, M. C., Jerrett, M., Pope III, C. A., Krewski, D., Gapstur, S. M., Diver, W. R., Beckerman, B. S., Marshall, J. D., Su, J., Crouse, D. L., and Burnett, R. T.: Long-term ozone exposure and mortality in a large prospective study, Am. J. Respir. Crit. Care Med., 193, 1134–1142, https://doi.org/10.1164/rccm.201508-1633OC, 2016.

US Environmental Protection Agency: Integrated science assessment for ozone and related photochemical oxidants, Office of Research and Development, Research Triangle Park, NC (February), EPA/600/R-10/076F, 2013.

Voulgarakis, A., Naik, V., Lamarque, J.-F., Shindell, D. T., Young, P. J., Prather, M. J., Wild, O., Field, R. D., Bergmann, D., Cameron-Smith, P., Cionni, I., Collins, W. J., Dalsøren, S. B., Doherty, R. M., Eyring, V., Faluvegi, G., Folberth, G. A., Horowitz, L. W., Josse, B., MacKenzie, I. A., Nagashima, T., Plummer, D. A., Righi, M., Rumbold, S. T., Stevenson, D. S., Strode, S. A., Sudo, K., Szopa, S., and Zeng, G.: Analysis of present day and future OH and methane lifetime in the ACCMIP simulations, Atmos. Chem. Phys., 13, 2563–2587, https://doi.org/10.5194/acp-13-2563-2013, 2013.

Watanabe, S., Hajima, T., Sudo, K., Nagashima, T., Takemura, T., Okajima, H., Nozawa, T., Kawase, H., Abe, M., Yokohata, T., Ise, T., Sato, H., Kato, E., Takata, K., Emori, S., , and Kawamiya, M.: MIROC-ESM 2010: Model description and basic results of CMIP5-20c3m experiments, Geosci. Model Dev., 4, 845–872, https://doi.org/10.5194/gmd-4-845-2011, 2011.

Weatherhead, E. C., Bodeker, G. E., Fassò, A., Chang, K.-L., Lazo, J. K., Clack, C. T. M., Hurst, D. F., Hassler, B., English, J. M., and Yorgun, S.: Spatial coverage of monitoring networks: A climate observing system simulation experiment, J. Appl. Meteorol. Climatol., 56, 3211–3228, https://doi.org/10.1175/JAMC-D-17-0040.1, 2017.

Weigel, A. P., Knutti, R., Liniger, M. A., and Appenzeller, C.: Risks of model weighting in multimodel climate projections, J. Clim., 23, 4175–4191, https://doi.org/10.1175/2010JCLI3594.1, 2010.

Williamson, D., Blaker, A. T., Hampton, C., and Salter, J.: Identifying and removing structural biases in climate models with history matching, Clim. Dyn., 45, 1299–1324, https://doi.org/10.1007/s00382-014-2378-z, 2015.

Wood, S. N.: Generalized additive models: an introduction with R (2nd Edition), CRC press, 2017.

Wood, S. N. and Augustin, N. H.: GAMs with integrated model selection using penalized regression splines and applications to environmental modelling, Ecol. Model., 157, 157–177, 2002.

Wood, S. N., Bravington, M. V., and Hedley, S. L.: Soap film smoothing, J. Roy. Stat. Soc. B, 70, 931–955, https://doi.org/10.1111/j.1467-9868.2008.00665.x, 2008.

World Health Organization: Air quality guidelines global update 2005: Particulate matter, ozone, nitrogen dioxide, and sulfur dioxide, WHO, Regional Office for Europe, Copenhagen, http://www.euro.who.int/__data/assets/pdf_file/0005/78638/E90038.pdf, 2005.

World Meteorological Organization: Scientific Assessment of Ozone Depletion: 2010: Pursuant to Article 6 of the Montreal Protocol on Substances that Deplete the Ozone Layer, World Meterological Organization, 2011.

Wu, S., Mickley, L. J., Jacob, D. J., Rind, D., and Streets, D. G.: Effects of 2000–2050 changes in climate and emissions on global tropospheric ozone and the policy-relevant background surface ozone in the United States, J. Geophys. Res. Atmos., 113, https://doi.org/10.1029/2007JD009639, 2008.

Young, P. J., Archibald, A. T., Bowman, K. W., Lamarque, J.-F., Naik, V., Stevenson, D. S., Tilmes, S., Voulgarakis, A., Wild, O., Bergmann, D., Cameron-Smith, P., Cionni, I., Collins, W. J., Dalsøren, S. B., Doherty, R. M., Eyring, V., Faluvegi, G., Horowitz, L. W., Josse, B., Lee, Y. H., MacKenzie, I. A., Nagashima, T., Plummer, D. A., Righi, M., Rumbold, S. T., Skeie, R. B., Shindell, D. T., Strode, S. A., Sudo, K., Szopa, S., and Zeng, G.: Pre-industrial to end 21st century projections of tropospheric ozone from the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP), Atmos. Chem. Phys., 13, 2063–2090, https://doi.org/10.5194/acp-13-2063-2013, 2013.

Young, P. J., Naik, V., Fiore, A. M., Gaudel, A., Guo, J., Lin, M., Neu, J. L., Parrish, D. D., Rieder, H. E., Schnell, J. L., Tilmes, S., Wild, O., Zhang, L., Ziemke, J. R., 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., and Zeng, G.: Tropospheric Ozone Assessment Report: Assessment of global-scale model performance for global and regional ozone distributions, variability, and trends, Elem. Sci. Anth., 6, https://doi.org/10.1525/elementa.265, 2018.

Table 1. List of the ensemble members used in this paper.

Model ID	Group	Resolution	Meteorological Forcing[†]	References
CHASER (MIROC-ESM)	Nagoya University; Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Japan	2.8° × 2.8°	C2	Sudo et al. (2002a, b); Watanabe et al. (2011)
GEOSCCM	NASA Goddard Space Flight Center, USA	2.5° × 2°	C2	Oman et al. (2011)
GFDL-AM3	NOAA Geophysical Fluid Dynamics Laboratory, USA	2° × 2°	C1SD	Lin et al. (2012, 2014, 2017)
G5NR-Chem	NASA Goddard Space Flight Center, USA	0.125° × 0.125°	$*$	Hu et al. (2018)
MOCAGE	Centre National de Recherches Météorologiques; Météo France, France	2° × 2°	C2	Josse et al. (2004); Teyssèdre et al. (2007)
MRI-ESM1r1	Meteorological Research Institute, Japan	2.8° × 2.8°	C2	Adachi et al. (2013)

† Meteorological forcing includes coupled ocean-atmosphere (C2) and nudged to observed reanalysis meteorology (C1SD) in CCMI reference simulations (Morgenstern et al., 2017).

$*$ The specification of forcing scenario for this special run is described by Hu et al. (2018).

Table 2. RMSEs (averaged errors in a given region) between spatially interpolated observations and each model, along with regionally optimized weights $\{\beta_{rk}$: for k-th model in region r$\}$ (zero weights are not displayed). Last column shows the RMSEs from equal weighted averages or constrained weights from the multi-model composite. All the numbers are reported in units of ppb (i.e. parts per billion by volume).

				Regional RMSE				Averaged
Region		CHASER	GEOSCCM	GFDL-AM3	G5NR-Chem	MOCAGE	MRI-ESM1r1	Error
Africa		6.40	8.91	12.16	12.16	10.47	14.89	10.83
N America		10.04	8.90	11.28	9.20	24.39	8.41	12.04
S America		7.39	7.19	10.00	8.81	10.59	8.59	8.76
E Asia		9.12	9.42	15.89	13.33	17.68	14.40	13.31
S/C Asia		7.68	15.11	13.36	13.38	13.37	18.41	13.55
Europe		9.14	8.41	10.75	8.20	11.88	9.66	9.67
Oceania		6.00	6.81	11.82	9.42	9.38	9.24	8.78
Russia		6.59	9.10	11.71	7.86	20.29	6.04	10.27
				Constrained weights of the multi-model composite				Composite
Region	α_r	CHASER	GEOSCCM	GFDL-AM3	G5NR-Chem	MOCAGE	MRI-ESM1r1	Error
Africa	-5.25	0.27	0.12	0.43	0.01	0.17	-	5.39
N America	-7.84	-	0.38	-	0.62	-	-	4.35
S America	2.13	0.63	0.13	-	0.24	-	-	5.37
E Asia	-7.99	0.08	0.83	0.09	-	-	-	4.88
S/C Asia	-8.90	0.52	0.26	0.12	0.10	-	-	4.95
Europe	-9.91	-	-	0.78	0.13	0.09	-	2.75
Oceania	-2.36	0.73	-	-	0.27	-	-	5.76
Russia	-7.15	0.21	-	0.45	0.32	0.02	-	2.04

Table 3. RMSE against TOAR observations (i.e. not interpolated ozone) from the multi-model mean (MMM), multi-model composite (from fusion step 2) and the final fused product (from fusion step 3).

	MMM	Composite	Fusion
E Asia	14.44	5.72	4.27
Europe	11.64	5.31	4.26
N America	12.22	4.51	2.76
Overall$*$	12.32	5.16	3.82

$*$ Overall category includes all available sites around the world.

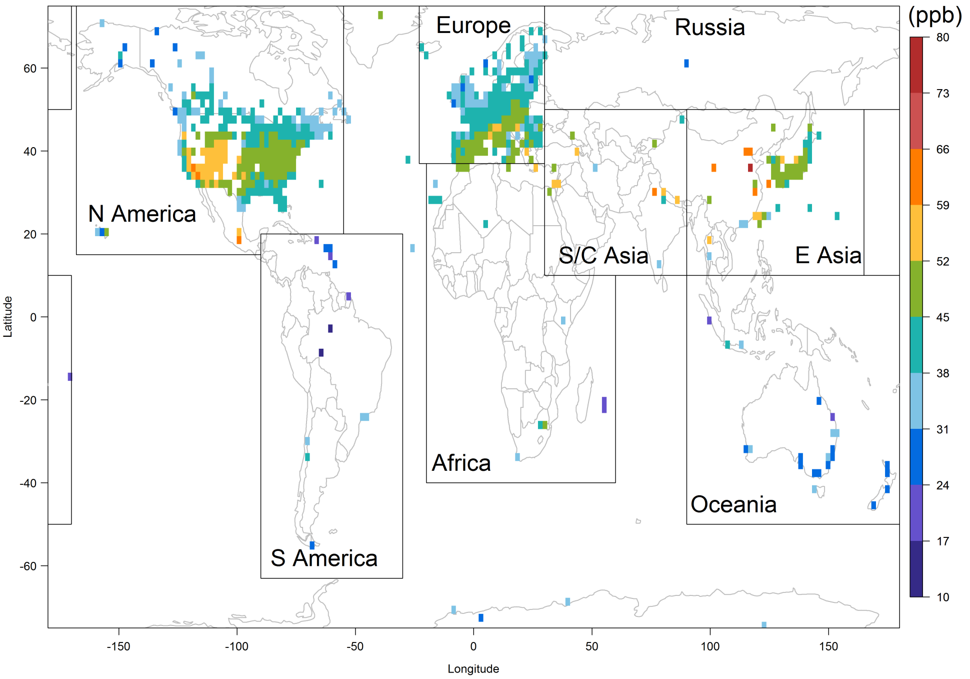
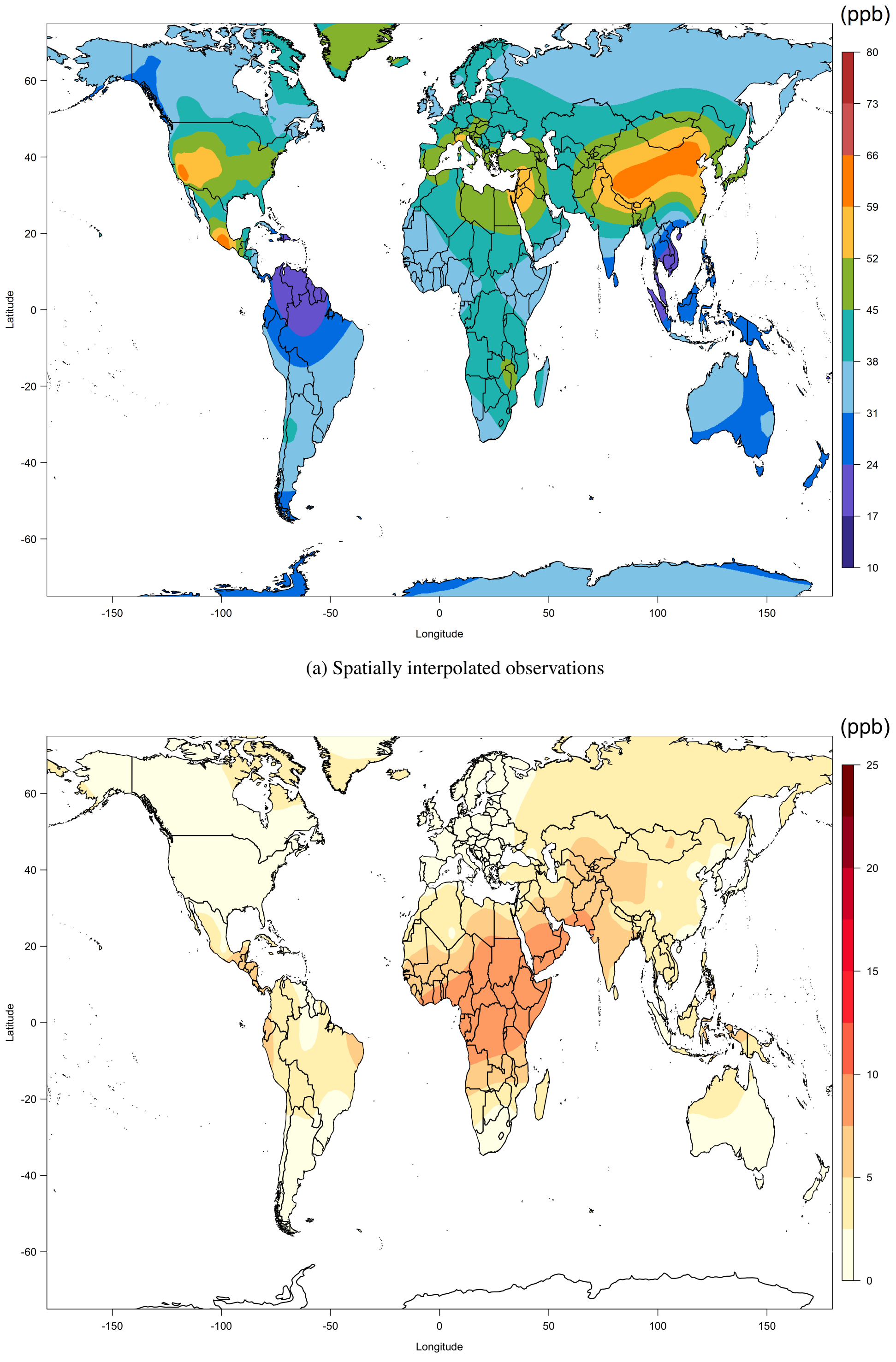


Figure 1. TOAR observations where the monitoring locations are discretized to a 2° × 2° grid in 2008-2014.

Table A1. Summary of results from fitting nine candidate statistical models (annual average over 2008-2014).

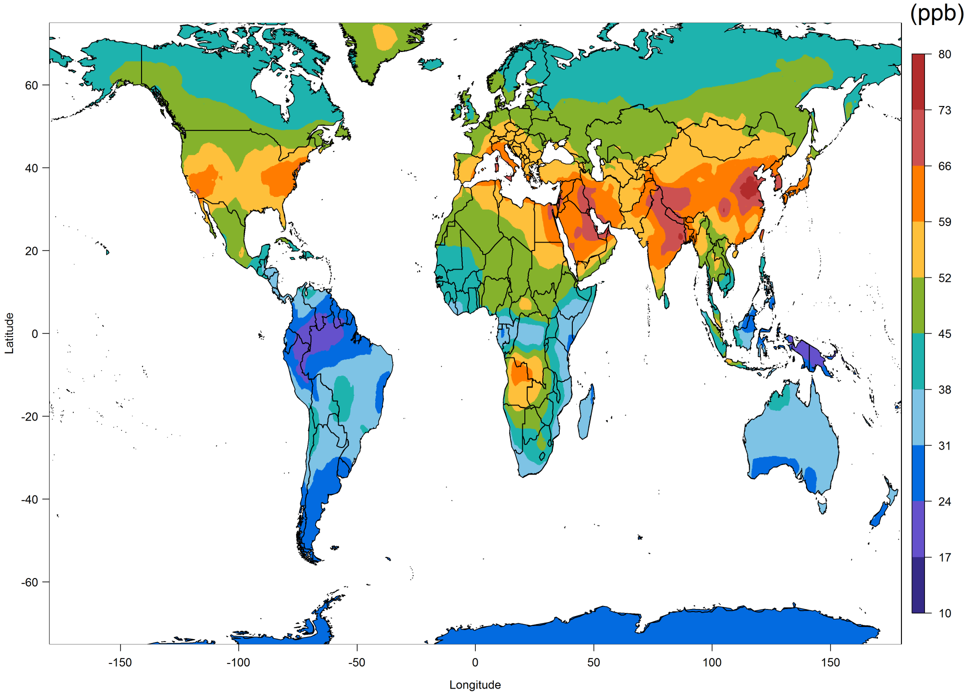
# basis	0	1	2	3	4	5	6	7	8
RMSE	3.82	3.17	3.18	3.23	2.90	2.52	2.76	2.76	3.44
DIC	-1517	-1548	-1556	-1561	-1593	-1621	-1603	-1594	-1565
GCV	2.78	2.64	2.62	2.60	2.50	2.43	2.44	2.48	2.60



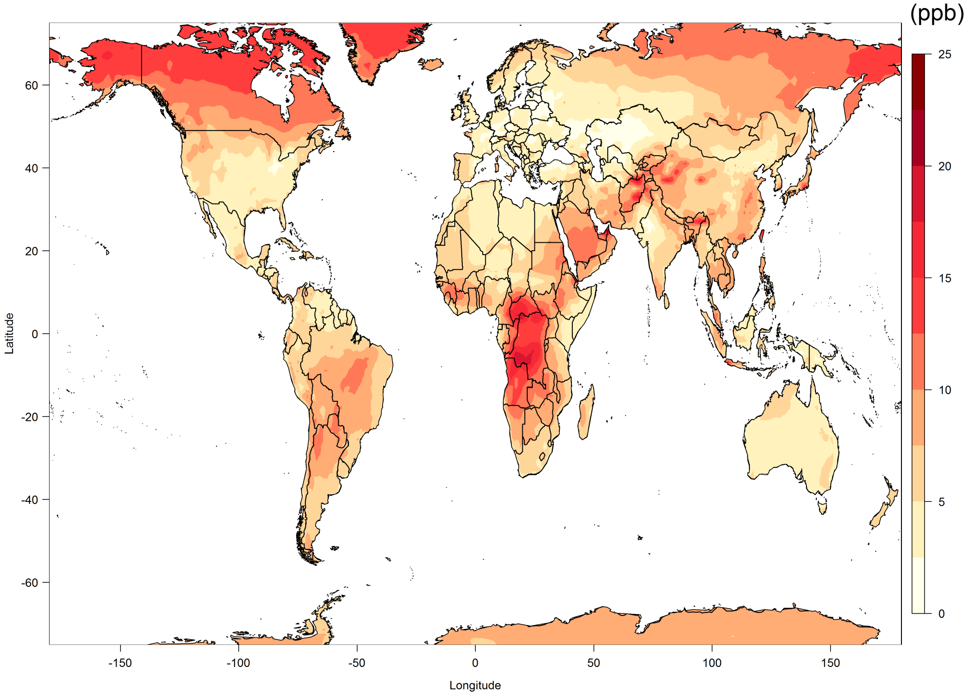
(a) Spatially interpolated observations

(b) Interpolation uncertainty

Figure 2. Estimates of spatially interpolated surface ozone distribution and associated uncertainty (half-width of the 95% credible interval from each cell).

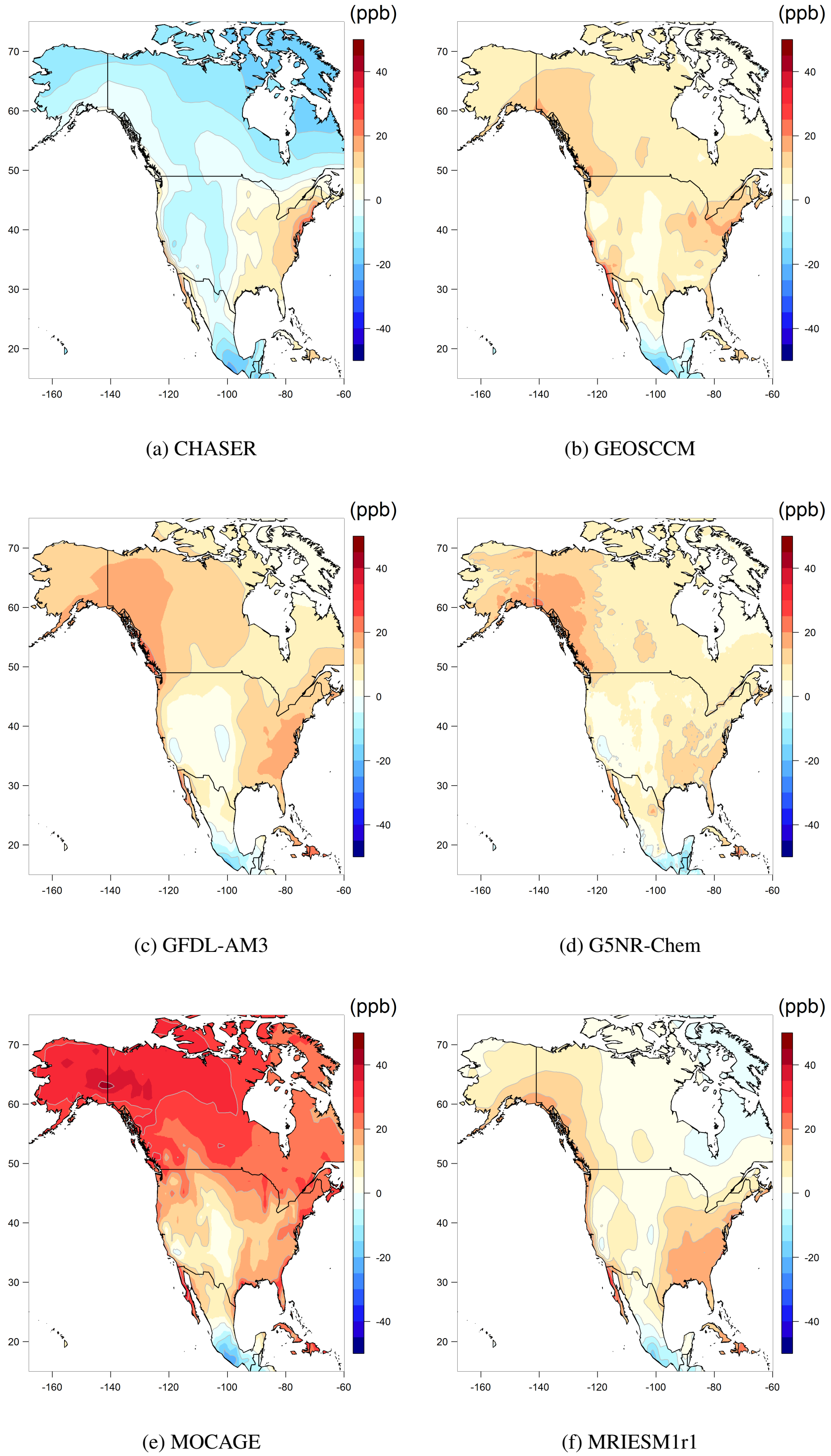


(a) Cell-by-cell mean



(b) Cell-by-cell SD

Figure 3. Multi-model mean and standard deviation (SD) in each grid cell from 6 ensemble members.



(a) CHASER

(b) GEOSCCM

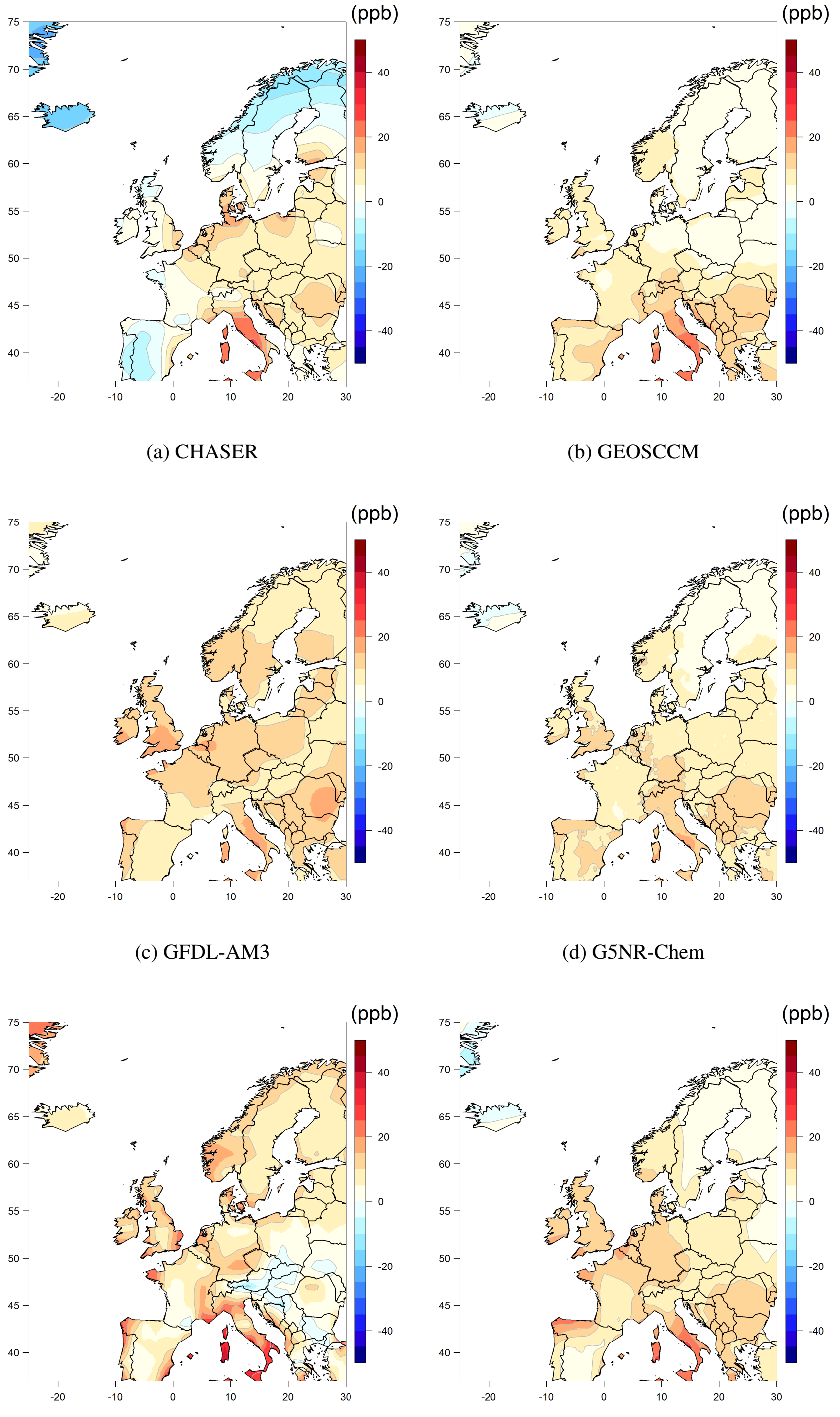
(c) GFDL-AM3

(d) G5NR-Chem

(e) MOCAGE

(f) MRIESM1r1

Figure 4. Spatial distributions of the ozone metric in North America from each model minus spatially interpolated observations.



(a) CHASER

(b) GEOSCCM

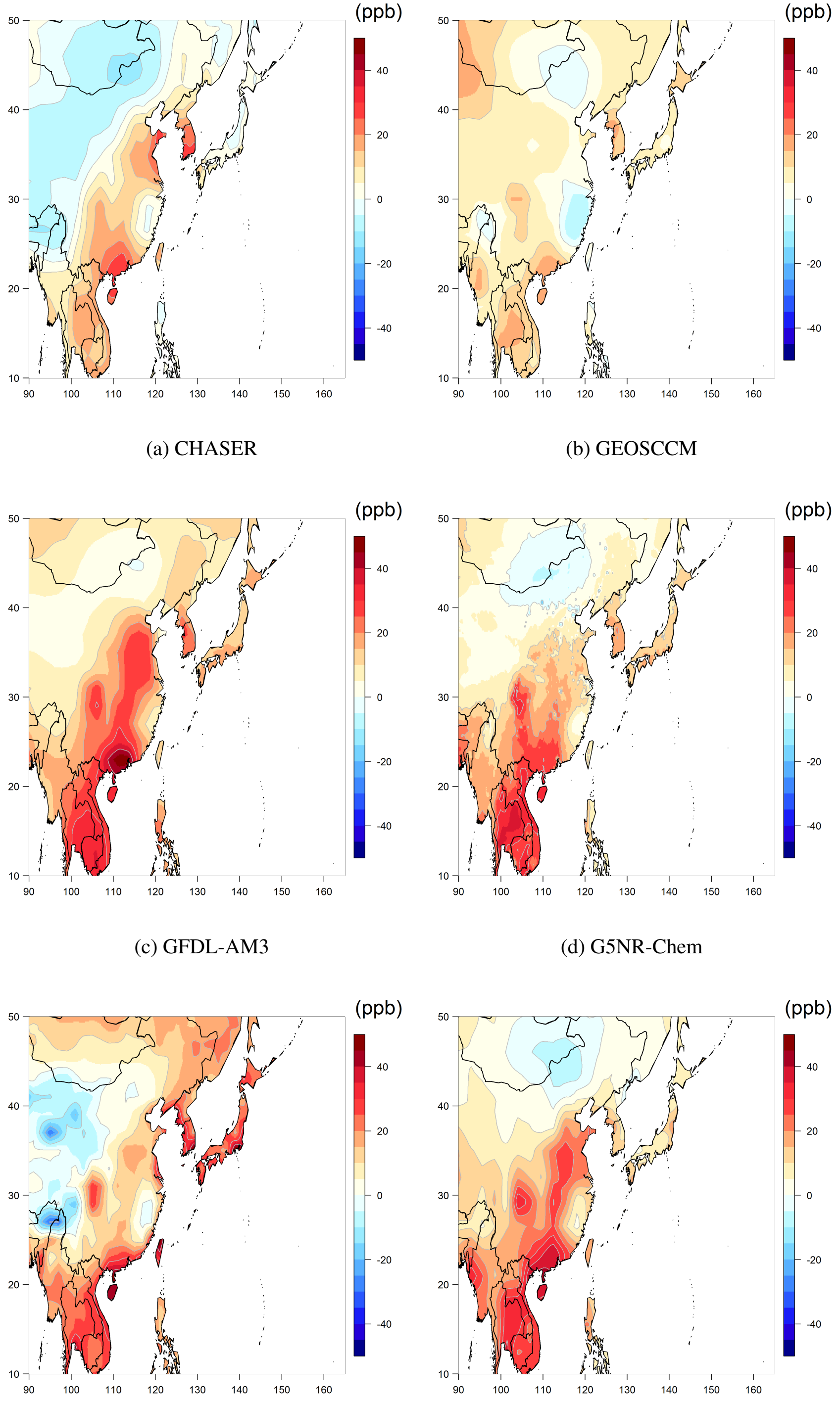
(c) GFDL-AM3

(d) G5NR-Chem

(e) MOCAGE

(f) MRIESM1r1

Figure 5. Spatial distributions of the ozone metric in Europe from each model minus spatially interpolated observations.



(a) CHASER

(b) GEOSCCM

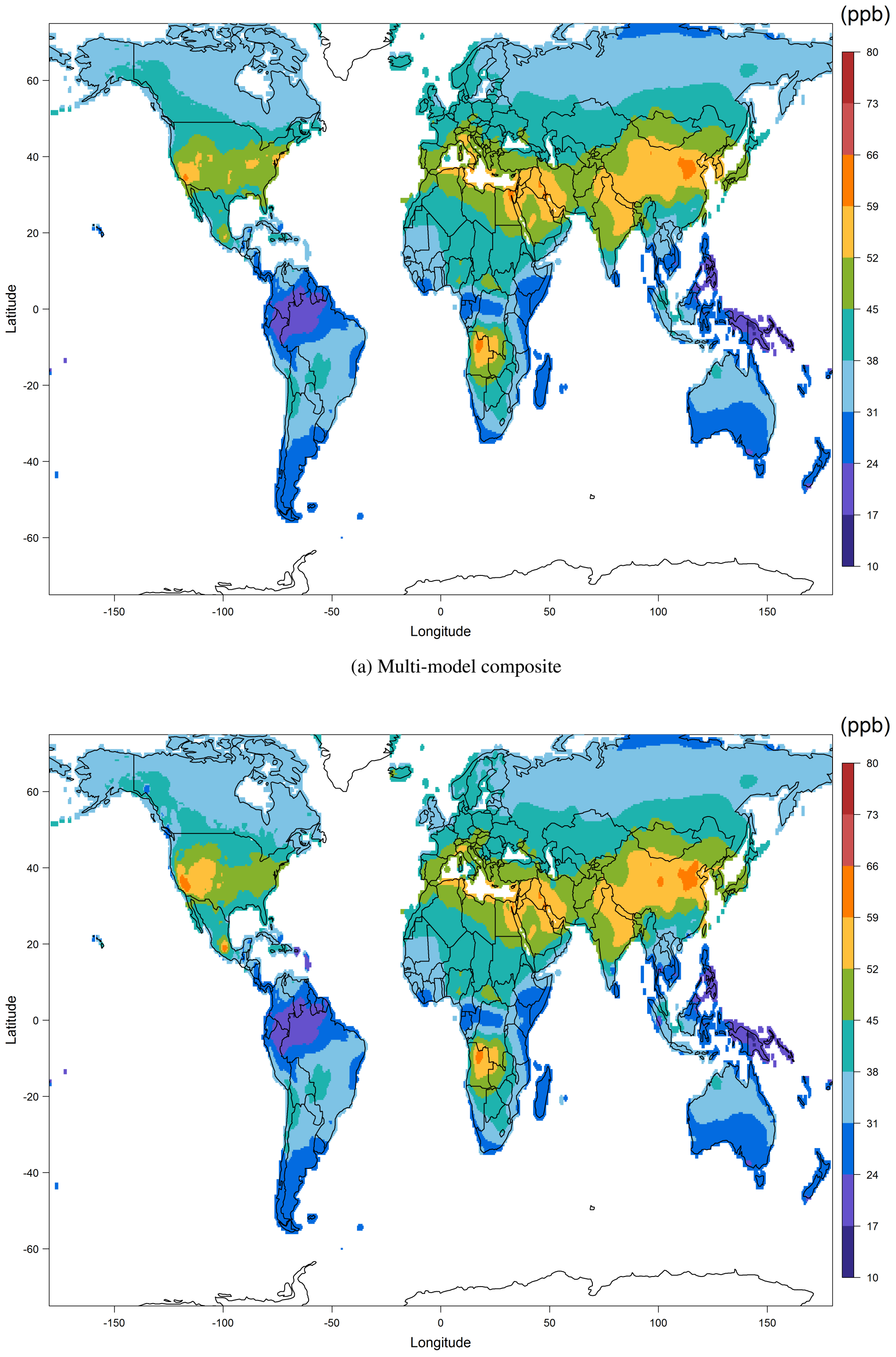
(c) GFDL-AM3

(d) G5NR-Chem

(e) MOCAGE

(f) MRIESM1r1

Figure 6. Spatial distributions of the ozone metric in East Asia from each model minus spatially interpolated observations.



(a) Multi-model composite

(b) Multi-model composite + bias correction

Figure 7. Multi-model composite and bias corrected surface.

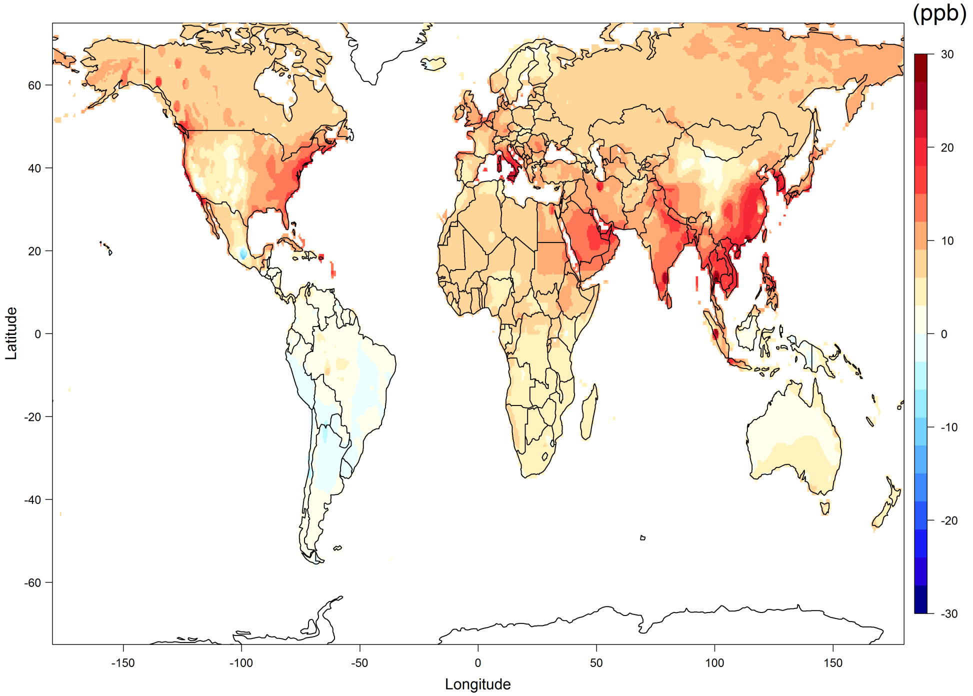


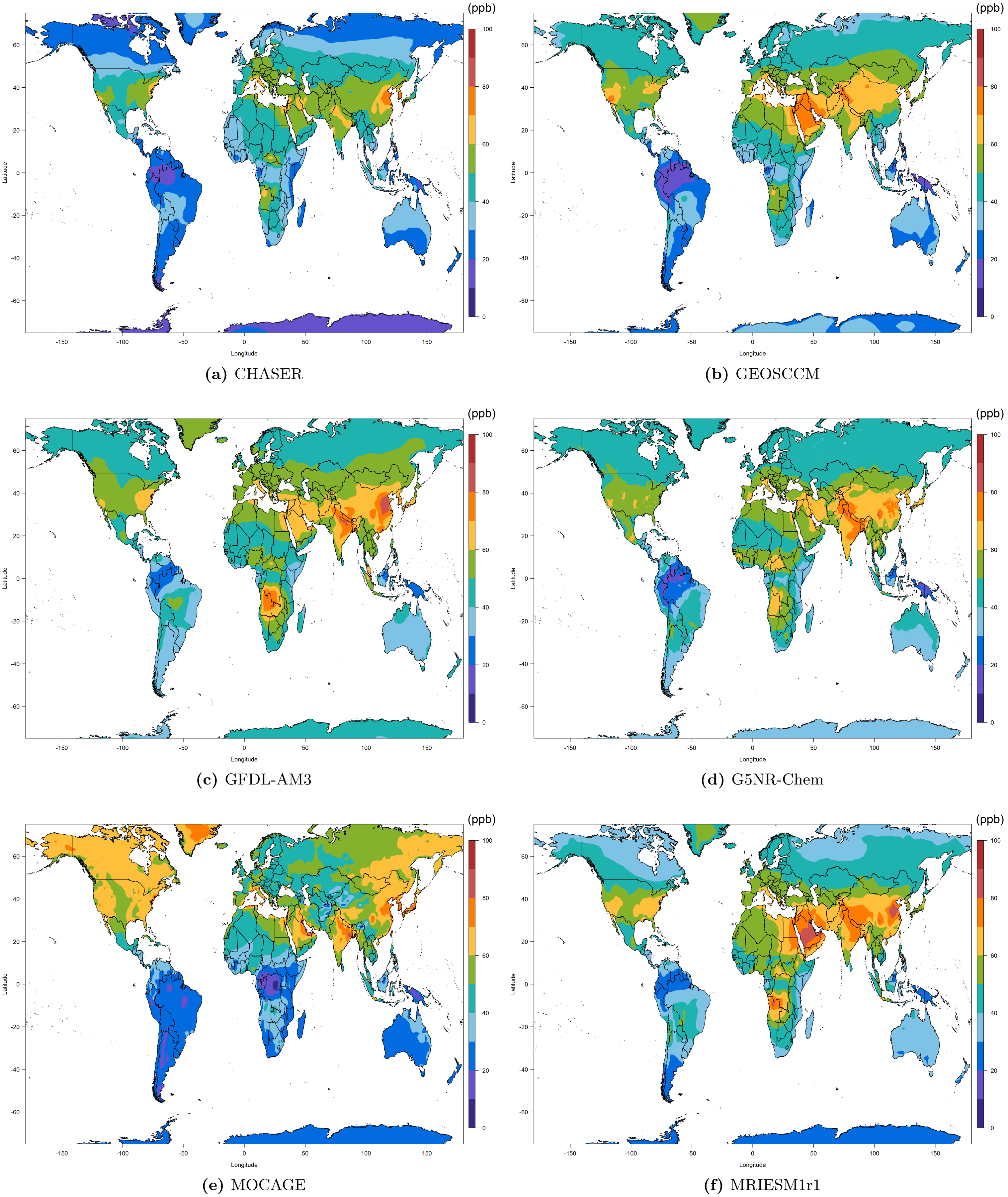
Figure 8. Map showing result for multi-model mean minus the fused surface ozone.

Supplemental Material for A new method (M^3Fusion-v1) for combining observations and multiple model output for an improved estimate of the global surface ozone distribution

Kai-Lan Chang[1, 2, 3, *], Owen R. Cooper[2, 3], J. Jason West[4], Marc L. Serre[4], Martin G. Schultz[5], Meiyun Lin[6, 7], Virginie Marécal[8], Béatrice Josse[8], Makoto Deushi[9], Kengo Sudo[10, 11], Junhua Liu[12, 13] and Christoph A. Keller[12, 13, 14]

[1]National Research Council Fellow
[2]NOAA Earth System Research Laboratory, Boulder, CO, USA
[3]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA
[4]Department of Environmental Sciences & Engineering, University of North Carolina, Chapel Hill, NC, USA
[5]Jülich Supercomputing Centre (JSC), Forschungszentrum Jülich, Jülich, Germany
[6]NOAA Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA
[7]Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ, USA
[8]Météo-France, Centre National de Recherches Météorologiques, Toulouse, France
[9]Meteorological Research Institute (MRI), Tsukuba, Japan
[10]Graduate School of Environmental Studies, Nagoya University, Nagoya, Japan
[11]Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokosuka, Japan
[12]NASA Goddard Space Flight Center, Greenbelt, MD, USA
[13]Universities Space Research Association, Columbia, MD, USA
[14]John A. Paulson School of Engineering and Applied Science, Harvard University, Cambridge, MA, USA
*Corresponding author: kai-lan.chang@noaa.gov

List of Figures



(a) CHASER

(b) GEOSCCM

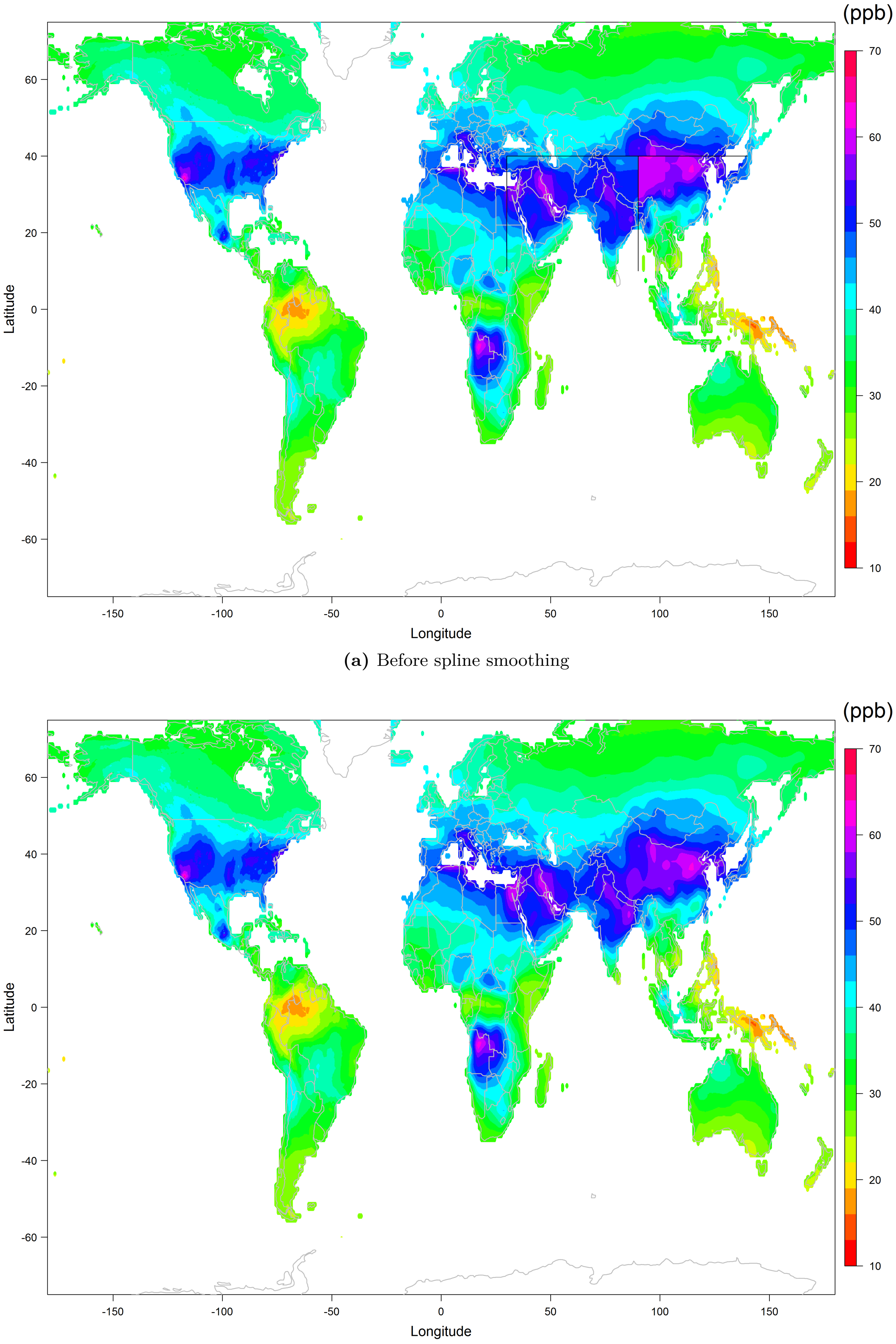
(c) GFDL-AM3

(d) G5NR-Chem

(e) MOCAGE

(f) MRIESM1r1

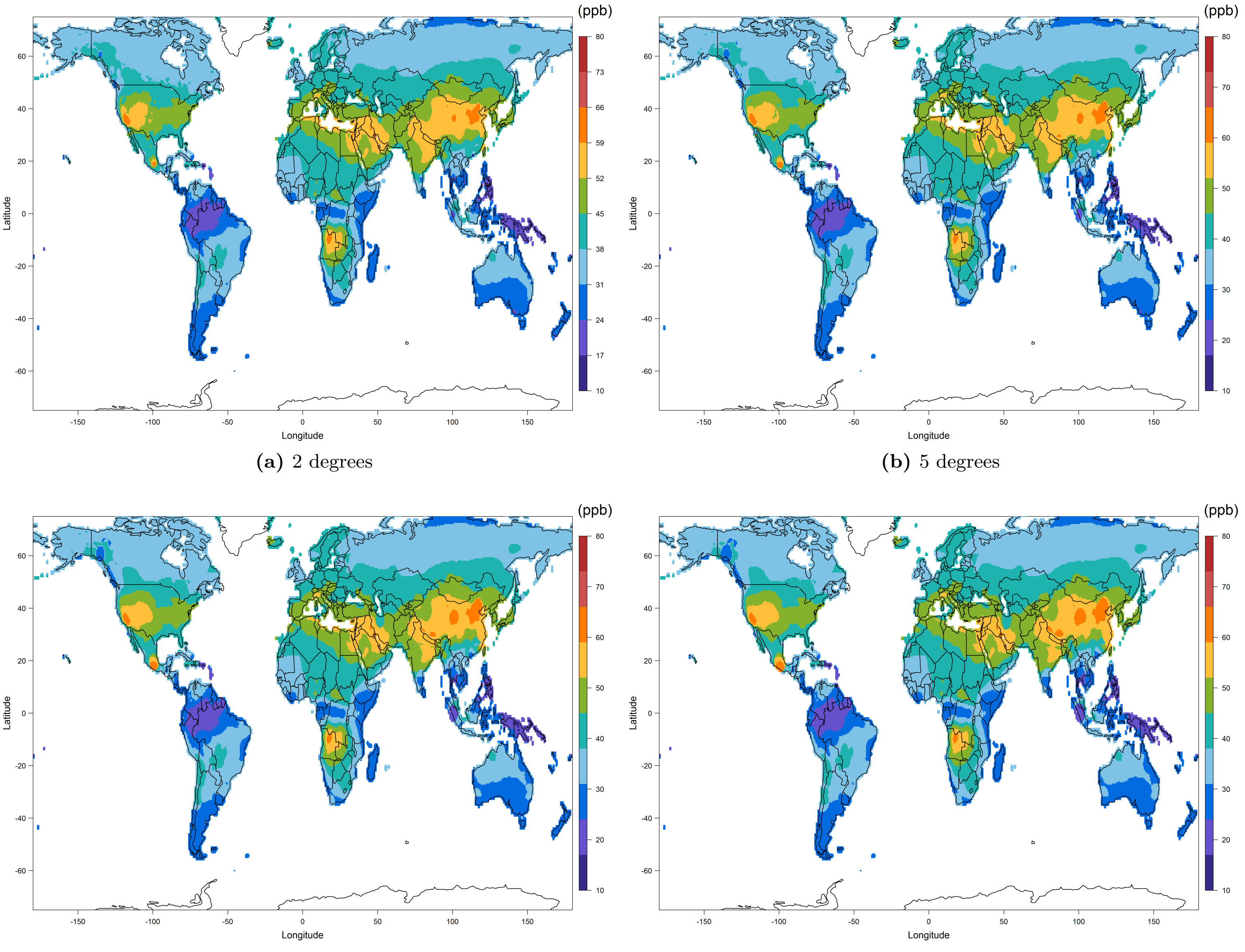
Figure S-1: Global distributions of the ozone metric from ensemble members.



(a) Before spline smoothing

(b) After spline smoothing

Figure S-2: Strong ozone discontinuities, or artefacts, were present along the geometric boundaries, especially in western China, before a spline smoothing was employed. The smoothing is only applied to 3 regions: one horizontal discontinuity between Russia and East/South Asia, one vertical discontinuity between East and South Asia, and one vertical discontinuity between South Asia and Africa.



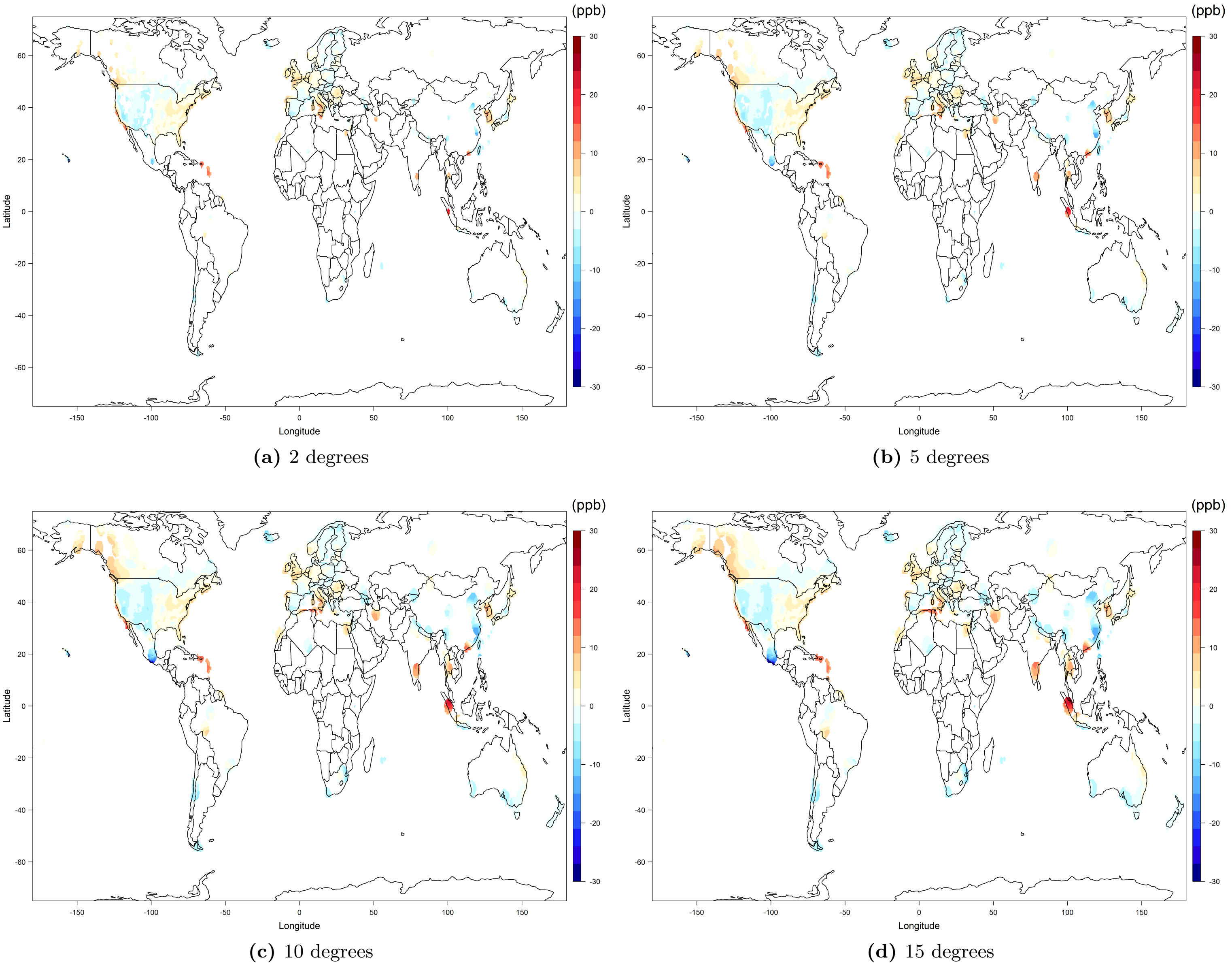
(a) 2 degrees

(b) 5 degrees

(c) 10 degrees

(d) 15 degrees

Figure S-3: The multi-model bias corrected surface under different ranges of correction radius.



(a) 2 degrees

(b) 5 degrees

(c) 10 degrees

(d) 15 degrees

Figure S-4: Amplitudes of multi-model bias correction under different ranges of correction radius.

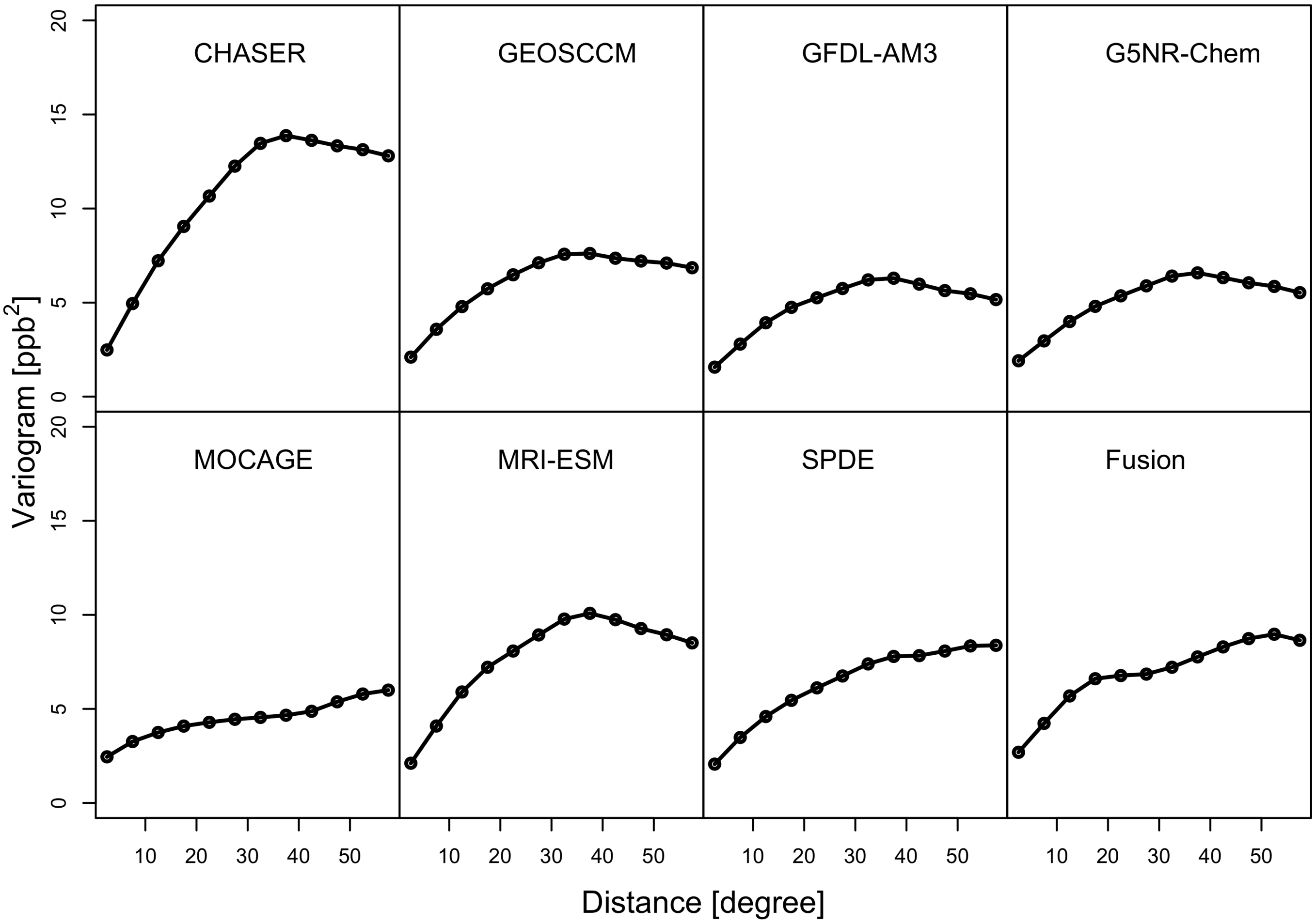


Figure S-5: The empirical variogram of ozone metric in North America from each model and product.