Response to Reviewers' Comments on "Radiative forcing and climate metrics for ozone precursor emissions: the impact of multi-model averaging" by C. R. MacIntosh et al.

www.atmos-chem-phys-discuss.net/14/27195/2014/

We would like to thank both referees for their constructive comments. The referees' suggestions have greatly improved the manuscript. In this response, each comment is addressed in turn, with responses given in bold-italic face and extracts from the proposed revised text in bold face.

Some sections of the manuscript have undergone significant revisions in response to reviewer comments. These proposed revisions are provided as a supplement at the end of this document, listed in the order in which they appear in the original manuscript, with the relevant page and line numbers . Each significant revision is numbered SR.x, and where one of the referees comments has been addressed with a significant revision, the relevant SR.x is referenced.

Reviewer 1

The authors examine how multi-model ensemble averaging impacts the calculation of radiative forcing, GWP, and GTP from changes in ozone and its precursors, and the related uncertainties. They conclude that using the multi-model mean as input to the radiative forcing code makes no significant difference relative to using averaged results from individual ensemble members. However there are significant differences in the estimated uncertainties between the two approaches; this should be taken into account when assessing the uncertainties associated with simply using the ensemble means as input, and the uncertainties are larger than the true uncertainties from calculating the radiative forcing from individual models. This is a valid topic worth publishing.

We thank the reviewer for these positive comments, and that the reviewer finds the topic worth publishing

However, I feel that the presentation can be improved, and the paper should be carefully revised to make it more readable and to focus on what the authors really want to deliver. In the present form, there is too much technical jargon and too many details that obscure the main points. There are many places in the text that need to be clarified.

We accept these criticisms – if the paper is not emphasizing our key message, then we needed to rewrite to ensure that it does.

Section 2 ("Methods") should be re-organized to more clearly describe what methods you use, and should give brief descriptions also of the established methodologies, e.g. Fry et al. At present, these messages are very obscure and are everywhere; this makes the text difficult to follow. This section could also be condensed to just present the essential message.

We have condensed the methods section to present the essential message. Significant manuscript revisions which address the methods section are presented as SR.2 in the supplement to this response.

A clear message on why using ensemble means result in larger uncertainties would be useful. Sections 4 and 5 should be condensed to deliver the main points/messages more clearly. Presently it is very difficult to follow all these details.

We have condensed Sections 4 and 5 so that they concentrate on our main message (the impact of averaging of model ensembles) rather than the values of the RF/GWP/GTP, which has already been discussed in the literature. The focus of the introduction has also been shifted to better frame the work. Significant manuscript revisions of the introduction are presented in SR.1, of the Section 4 discussion in SR.3 and of Section 5 in SR.4

Specific comments:

1) The authors should state specifically what causes the discrepancies w.r.t. estimating the uncertainties using the alternative approach (i.e. the ensemble means) in the abstract and/or in the conclusion.

We have added text to make this clearer the specific cause of the discrepancies, although we note Reviewer 2's very positive view of the conclusions. The new text in the abstract reads:

However, estimates of the standard deviation calculated from the ensemble-mean input fields (where the standard deviation at each point on the model grid is added to or subtracted from the mean field) are almost always substantially larger in RF, GWP and GTP metrics than the true standard deviation, and can be larger than the model range for short-lived ozone RF, and for the 20 and 100 year GWP and 100 year GTP. The order of averaging has most impact on the metrics for NO_X, as the net values for these quantities is the residual of the sum of terms of opposing signs. For example, the standard deviation for the 20 year GWP is two to three times larger using the ensemble-mean fields than using the individual models to calculate the RF. The source of this effect is largely due to the construction of the input ozone fields, which overestimate the true ensemble spread.

Page 27196

2) Line 7-8: Do you mean methane concentration or just methane lifetime as the input to the RF code? The radiation code does not directly accept "lifetimes" as input.

We agree this was unclear. We meant that concentration changes were input to the RF. The new text now reads:

Multi-model ensembles are frequently used to assess understanding of the response of ozone and methane lifetime to changes in emissions of ozone precursors such as NOx, VOC and CO. When these ozone changes are used to calculate radiative forcing (RF) (and climate metrics such as the global warming potential (GWP) and global temperature potential (GTP)) there is a methodological choice, determined partly by the available computing resources, as to whether the mean ozone (and methane) concentration changes are input to the radiation code, or whether each model's ozone and methane changes are used as input, with the average RF computed from the individual model RFs.

3) Line 15-16: Please quantify how significant these numbers are in the overall estimation of the RF, GWP, and GTP (e.g. give absolute values).

We do not agree with this comment. The absolute numbers for, for example, the NO$_x$ GWP, would only have a useful context for comparison with, for example, the VOC GWP, if it was stated what the emissions of these gases are. The key message of this paper is the impact of averaging, which we think is best communicated with percentages.

4) Line 19: Spell out "SD" here.

The confusing usages of SD and sigma, which largely arose at the typesetting stage, has been clarified throughout, and following a comment by reviewer 2, we have also clarified how these quantities were calculated.

5) Line 23-24: "We find that the effect is generally most marked for the case of NOx emissions": What is the cause of this effect?

We agree that this was unclear. We have changed this sentence to read

The order of averaging has most impact on the metrics for NOx as the net values for these quantities is the residual of the sum of terms of opposing signs.

Page 27197

(6) Line 27: Please explain "primary mode" here, supported by a reference.

Following the comments of reviewer 2, we have removed all reference to the primary mode, and refer, instead, the short- and long-lived ozone responses.

Page 27198

7) Line 15: Please define "±σ" here.

See response to comment 4 above.

8) Line 25-28: This should be more specific. Please state clearly and in context what you are going to address in the following sections.

We have expanded this paragraph to be much clearer on the context of each section. The new text reads:

Section 2 introduces the HTAP data and scenarios, and describes the radiation code used to perform the radiative transfer calculations. The method of Fry et al. (2012) to generate the subset of fields for input to the radiation code is briefly described, together with a description of further preparing this output for generation of the GWP and GTP metrics. Section 3 presents the initial ozone and methane fields that serve as input to the radiation code for both methodologies, and briefly discusses their differences. Sections 4 and 5 discuss the effect of the different methodologies on the reported RF and GWP and GTP respectively, and conclusions are given in Section 6.

Page 27201

9) Line 13: “This approach differs . . .” – Can you summarise more clearly what your approach is and what exactly is the difference w.r.t. Fry et al. What is the simple formula from Ramaswamy et al. (2001)? Please explain/define “back-calculation”.

The text has been altered so that the formula is presented and we no longer use the term “back-calculation” to describe the process. Further, the reference to the IPCC report (Ramaswamy, 2001) has been replaced with the correct original reference, Myhre et al. (1998). The new text is part of a wider substantial re-working of the methods section, which is presented in SR.2, and reads:

This full model ensemble is contrasted with the method used in Fry et al. (2012). This method first constructs a representative subset of model input fields for input into the radiation code. This subset comprises the ensemble mean control fields, plus the ensemble mean ± standard deviation short-lived ozone, methane and long-lived ozone perturbations. This subset of fields is constructed as follows: Firstly, each model field for each month is regridded to a common resolution; secondly, the mean and standard deviation of the ozone field is calculated for each month, for each pixel at each level. The standard deviation is then added to or subtracted from the mean field to give a 3-D representative field for each month.

These fields are grouped into four cases; the first comprises the control fields; the second the mean total ozone change (i.e. the sum of the short- and long-lived mean ozone fields) together with the mean methane change; the third the mean plus standard deviation total ozone and methane change; and the final case the mean minus the standard deviation changes. Therefore the radiation code must run only three times for each HTAP scenario (plus once for the control run), relative to 33 (11 models, 3 gaseous species) plus 11 control runs for the complete case.

The subsetting method of calculation used in Fry et al. (2012) gives only the total RF for each scenario as output. The contributions to the total RF from each of the short-lived ozone, methane and long-lived ozone are then

calculated from this total. First, the methane RF is calculated from the change in concentration using the simple formula from Myhre et al. (1998)

$\Delta F - \alpha(\sqrt{M} - \sqrt{M0}) - (f(M,N0) - f(M0,N0))$ (1)

where $f(M,N) = 0.47\ln[1+2.01\times10-5(MN)0.75 + 5.31\times10-15M(MN)1.52]$, α is a constant, 0.12, N is N2O in ppb (constant at 315 ppb) and M is CH4 in ppb and the subscript 0 indicated the unperturbed case.

The difference between the total RF and this methane RF is then attributed to ozone. For the calculation of the GWP and GTP metrics, it is further necessary to separate the ozone RF between the short- and long-lived components. This is achieved by scaling the RF due to the (purely long-lived) ozone perturbations in the SR2 scenario by the ratio of the long-lived ozone change in any given scenario and the SR2 scenario. This RF is attributed to the long-lived ozone, with the final residual being attributed to the short-lived ozone. The mean and standard deviation of the RF calculated using this subset of fields are denoted RF_{EN}.

Page 27205

10) Line 18: “Confidence in the chemistry of each species can be inferred”: I cannot understand how exactly such confidence can be inferred from the following statements.

This statement has been removed, as it added nothing helpful to our discussion

Figures

11) Plots in Figures 1&2 are too small. Units are missing on left axis on Figures 5&6.

We have redrawn Figures 1 and 2 so they have less whitespace and the general shortening of the text means that Figures 5 and 6 are no longer included.

Reviewer 2

This discussion paper by Macintosh, Shine and Collins addresses an important and often overlooked problem in averaging model results, one that is becoming more and more embedded in our chemistry-climate assessments. There is some very interesting material in the paper, but it is so long and unfocused, the continuing strings of deltas or perturbations become confusing, and as well it misses some of the basics that the community has already been through. My view is that it needs to put this work in a better perspective of the known correlations and chemical reactivities of the atmosphere and then state clearly what is new here and what is important. Hopefully this can be achieved in fewer pages so that the less stout can find the important results. Further, the use of un-normalized NOx perturbations makes the results not useful as some of the model spread (but not most) is spurious.

We note and accept the overall criticisms of the reviewer but we are pleased that it is recognised that it contains very interesting material. We have extensively revised the text, and removed two figures to reduce its length, and to focus on key results. We noted in particular the very positive comments of the reviewer on our Section 6, and this provided an impetus to improve the framing of the paper throughout. Since most of the results for the GWP and GTP metrics are, by their very definition, normalised, the issue of un-normalised results only affected one part of the paper which we have now alleviated.

27197ff This intro spends much time on the values and issues of GWP/GTPs but that is not what the paper is about. It really is about averaging, how to average, and how correlated errors can cancel and reduce uncertainty (or at least the model spread). The paper seems to have missed already published discussion on this topic, even within the limited framework of NOx, O3, CH4 and climate forcing. For example, the Holmes et al 2011 PNAS paper ("Uncertainties in climate assessment for the case of aviation NO") clearly points out the correlation of model results for RF short-O3, long-O3 and longCH4 and then shows how it affects the model spread is smaller than the sum of the components (a conclusion here). It would be better to start from that framework and build on it here with the self-consistent calculation of RF from the 3D fields (as is done).

We agree that there was too much focus on actual values and issues, rather than focusing on the averaging issue. We have refocused the introduction. We have also included a short discussion of Holmes et al. as this is clearly relevant to what we present. The substantial revisions to the introduction are presented in section SR.1. The discussion of Holmes et al. reads:

In the present work, we calculate the RF, GWP and GTP using output from each individual model in the HTAP ensemble. We then compare our results to those obtained with the ensemble- mean method of Fry et al. (2012). Hence, we can quantitatively assess the extent to which the RF calculated with the mean fields accurately represents the mean of the RF calculated using the ozone fields from each model individually. Further, by comparing the estimates of model and metric uncertainty (as represented by the standard deviations) in RF, and in GWP and GTP, we can assess whether such a representative subset can be used to accurately convey the spread in derived climate metrics. The result of this assessment will then guide the extent to which the use of the computationally less expensive ensemble-mean fields can be used, without compromising the quality of information.

The particular case of NOx is interesting because cancellation between RF due to different components of the total RF (and hence the GWP and GTP) can substantially reduce model spread (Holmes et al., 2011), if individual components are correlated. Using values drawn from the aviation NOx literature, they found that in general, a large (positive) RF due to the short-lived ozone forcing (driven directly by the NOx) in any one model, was associated with a large (negative) long-lived ozone forcing (driven indirectly by

the effect of NOx on methane concentrations) in the same model. Hence the uncertainty in the net RF, derived from considering the uncertainty in each component on its own, was found to be almost double the uncertainty in the net RF when the correlation was taken into account. Our work builds on Holmes et al. (2011) by exploiting results from a single multi-model intercomparison, and investigating the effects of different timescales on the cancellation, for emissions from a number of different regions, and extends it to CO and VOC (where the cancellation present in the NOx case is not present).

27198ff While the Collins et al 2013 paper clearly showed the regional differences in emissions-to-impacts, the much earlier work of Wild et al 2001 GRL ("Indirect longterm global radiative cooling from NOx emissions") has a figure/table showing the clear cancellation of the RF of short-O3 with that of long-O3+CH4 as well the large differences in the absolute impacts according to the latitude of emissions. There is a clear disagreement between that paper and the results presented here (p. 27204 & Fig.3) that had me stumped until I read carefully and found that none of the results had been normalized to a standard perturbation (e.g., 1 Tg-N/yr) per region. Furthermore, the individual models perturbations were %s and not absolutes – all of these perturbations need to be renormalized to make sense, and further the perturbations for the 4 HTAP regions must also be rescaled. It makes no sense to argue that the SA impacts of NOx are small when the perturbation is much smaller – I could not find these key numbers in the tables.

We agree that in this case, normalised values are more useful, and more consistent with the discussion of climate metrics GWP and GTP, which require input which is normalised to emission mass changes. The substantial revisions to Section 4 are presented in SR.3. Table 3 and Figure 3 have also been updated to reflect this change.

27213-4 The discussion Section 6 is very good, and I began to realize the value of this work. It would be helpful if the authors focused from the beginning on what was new here, and how by using the HTAP runs, imperfect as they are, we can learn something about 'ensembling'.

We appreciate this comment and will carry over the format of Section 6 to the rest of the paper.

Various points. There are many confusing points in the paper as well as ambiguously defined calculations. I give a quick run-through of my notes below in the order they appear:

The use of +-(greek sigma) and SD are not clear. Is there a difference?

No there is no difference. The problem mostly arose at the type-setting stage, and we failed to spot it. This has been alleviated.

The definition of sigma/SD must also be clear as to what time/space series is being used (hourly, daily, monthly) and over what period the SD is computed. Resolution is also critical and finer resolution will always have greater variability.

We agree that this needed to be much clearer as in some cases. We have now clarified how the standard deviation of the ozone fields is calculated, to apply the Fry method (i.e. for each month, at each grid point and at each level, the SD is calculated using the data from each model). For the full method, the SD we present is calculated using annual and global mean RF/GWP/GTP from each of the models. Concerning the final sentence, since each model is regridded to a common resolution, we compare each model on a common basis.

The text in the methods section has been updated to read:

This full model ensemble is contrasted with the method used in Fry et al. (2012). This method first constructs a representative subset of model input fields for input into the radiation code. This subset comprises the ensemble mean control fields, plus the ensemble mean ± standard deviation short-lived ozone, methane and long-lived ozone perturbations. This subset of fields is constructed as follows: Firstly, each model field for each month is regridded to a common resolution; secondly, the mean and standard deviation of the ozone field is calculated for each month, for each pixel at each level. The standard deviation is then added to or subtracted from the mean field to give a 3-D representative field for each month.

And in the description of Figure 2 it has been updated to read

The two sets of bars represent the spread in the model ensemble and denote the model standard deviation calculated in two different ways. Those in blue show the standard deviation calculated from the global-average burden change for each individual model. Those in red show the area-average of the 3D grid-point-level standard deviation fields, as in the subsetting method used by Fry et al. (2012). Here, the bars are calculated as the global annual-mean ± one standard deviation ozone field. The global average of the grid-point level standard deviation fields is not equal to the standard deviation calculated after the global mean for each model has been calculated, i.e. the order of operations in this case makes a substantial difference to the ± 1 standard deviation bars.

I believe that there 13 scenarios plus 1 control? = 14?

Agreed – thank you

You note that the 20% is not the same in all models, this should be rescaled.

Most of the values presented in the paper were already "per unit mass emission" as this is what is required for input to the metrics. We have clarified this. Substantial revisions to Section 4 (presented in SR.3), and updated versions of Table 3 and Figure 3 have brought this section in line with the rest of the paper.

The definition of tropopause when averaging is critical and I doubt you have hourly data, so you need to realize that a monthly or zonal mean tropopause height does not accurately separate the two regions. What was done here?

We have added text to make this clearer and to include caveats. Indeed we only have monthly data but we have calculated the tropopause locally on our common grid 2.75^{o}x3.75^o, rather than used zonal-means. The issue is certainly important, but we don't regard it as critical. New text reads:

The model output is re-gridded to a common resolution of 2.75o latitude x 3.75o longitude, with 24 vertical levels, which is comparable to the resolution of the models on average. A common tropopause was identified as the level at which the lapse rate falls below 2 K km−1. As many of the models do not include stratospheric chemistry, stratospheric changes in all species are neglected, and, above the tropopause, the models share a common climatology. Given the relatively coarse vertical resolution of the models, and that the data are monthly mean, any definition of tropopause is necessarily imperfect; however, this method ensures clarity when averaging monthly mean fields to form ensemble means, and minimises the aliasing of stratospheric ozone into the troposphere as part of the averaging process.

If you think about it, 4 points during the year hardly resolve the annual cycle, but they do reasonably sample it.

We agree and we have changed the wording from "resolve" to "sample" The new text now reads

For each model, January, April, July and October are used as input to the code, in order to reduce run-time constraints whilst remaining sufficient to reasonably sample the annual cycle in transport and RF.

The term uncertainty (p.27201) keeps slipping in when you mean model spread.

We agree and we have now made it clear when we mean model spread (which is indeed what we mean on most occasions).

Use of the abbreviation PM for primary-mode O3 or long-O3 is very odd and confusing. Finally in the conclusions you revert to the more standard short-O3 and long-O3 that is more standard. Primary mode is OK, but not PM. Because then the short-O3 should be Secondary Modes (plural).

We have acted on this comment and no longer refer to Primary Mode or PM. Although we accept the reviewers PM/SM logic, we note that primary mode in this context is in wide usage in this literature, and we see no difficulty in using PM as an abbreviation for Primary Mode (again it is widely done).

The discussion about calculating the steady-state CH4 abundance from the feedback factor is based on some very careful definitions of lifetimes, time scales and feedbacks etc in the literature – see the recent WMO and IPCC sections on this. The lifetime of CH4 must be defined to include ALL losses, otherwise the method here does not work.

It is unclear in Table 2 just how these "lifetimes" for CH4 are derived and thus how someone might usefully follow the chain of mapping the dln(lifetime)/dln(burden) onto a perturbation lifetime.

The methane lifetimes, and change in methane lifetimes are derived in Collins et al. (2013), not in this work. The text has been changed to make this clear, and to provide a brief summary of their method. The updated text (part of SR.2) reads :

The CTMs produce [OH],[O3] and associated atmospheric loss rates as 3-D output fields. Short-lived ozone can be used directly as input to the radiation code. Methane fields for each model and each simulation were globally homogeneous, and fixed at 1760 ppbv, except in the CH4 scenario, when they are reduced to 1408 ppbv. Equilibrium methane concentrations for each scenario have been calculated in Collins et al. (2013) from the change in methane lifetime, ∆α, as [CH4]=1760×((αcontrol+∆α)/ αcontrol)^f, where the methane lifetimes are calculated in Fiore et al. (2009) . These lifetimes include loss terms such as those to soil processes; however all those except the atmospheric term are assumed to be constant. The change in methane lifetime is also calculated in Collins et al. (2013) from the change in [OH] (since the atmospheric OH sink accounts for around 90% of loss of atmospheric CH4, and surface sinks are considered constant). Finally, the feedback factor, f is determined in Fiore et al. (2009) from the change in loss rates between the control and the CH4 perturbation scenarios, and accounts for the effect of methane change on its own lifetime (Prather, 1996). Further, long-lived changes also arise from the change in ozone resulting from a change in methane, which in turn depends on the change in methane lifetime for a given scenario. The long-lived ozone changes for each model and scenario are calculated as described in West et al. (2009) by scaling the ozone change in the CH4 perturbation simulation by the relative change in methane concentration in each scenario as given in Fiore et al. (2009).

I suspect that each model's VOC emissions are alos very different, not only in quantity, but also in their makeup. This will greatly increase the model spread for a 20% perturbation. Thus the arguments here about short lifetime (which sound plausible) may not be the reason.

This is a good point, thank you – we have changed the text to reflect the fact that there is an additional reason for inter-model differences for the VOCs. The updated text reads:

The largest standard deviations relative to the mean are found for the VOC case, in part due to large differences between the models in terms of VOC speciation and chemistry schemes (e.g. Collins et al. (2002)). Since each model defines its own VOC class within the chemistry scheme, the initial burden and the atmospheric lifetime can vary substantially between models.

In the Climate Metric section, there are so many numbers as to be confusing – some do not even have units (p. 27210).

We accept the basic criticism that the number of values made it difficult to read. We have depopulated the text with many of these, and let the tables speak for themselves. But any numbers we had presented were either dimensionless GWPs and GTPs (as these are all relative to CO_2) or else percentages, so we do not accept all this criticism. Significant revisions to this section are presented in SR.4.

Supplement detailing significant revisions to the manuscript.

SR.1 (p27197 L5-p27198 L28)
One method for characterising uncertainty in the climate sciences is to perform large, multi-model ensemble studies. This approach, provided that the range of models do indeed capture the range of climate responses to an applied perturbation, provides far more information, not only on the most likely climate response, but also on the likelihood of a range of possible responses - i.e. the uncertainty associated with the mean response. However, if further downstream analysis is performed on such a large model ensemble study, then methodological choices, which may be constrained by pragmatic concerns such as data processing time, must be made.

One common example of such an application of a model ensemble is in the calculation of climate metrics and their associated uncertainty. Climate metrics provide an important method of comparing the mean climate effects of emissions of various forcing agents. It is therefore desirable to be able to compute such metrics quickly and efficiently from input ensembles, but where possible without compromising on the quality of the reported values and, crucially, their associated measurements of model spread. Metrics such as the Global Warming Potential (GWP) and Global Temperature-change Potential (GTP, Shine et al. (2005)) introduce additional un- certainty and depend strongly on the time horizon, H that is under investigation,

but also on the spatial distribution of the forcing agent, and its lifetime in the atmosphere. These last two properties can vary strongly with model.

It would therefore seem reasonable to ask, what is the minimum volume of data processing and input information that can be used to provide meaningful estimates of climate metrics from large multi-model studies, without compromising the quality of the reported metrics and the representativeness of the associated spread.

The Hemispheric Transport of Air Pollutants (HTAP) study, provides a useful test case for the present work (Task Force on Hemispheric Transport of Air Pollution, 2010). A part of this project perturbed by emissions of species which are known to affect atmospheric ozone concentrations by 20% (in this case, NOx, VOC and CO). An ensemble of 11 chemistry transport models (CTMs) took part, and each perturbed the 3 precursors in 4 pre-defined source regions. Subsequent work by Fry et al. (2012) and Collins et al. (2013) assessed the RF, GWP and GTP for the precursor species. Computational limitations prevented the analysis of the variability in the RF, GWP and GTP in Fry et al. (2012); instead, the ensemble mean fields ± one standard deviation were deemed to provide the minimum subset of fields which could be used to represent the mean and standard deviation in the derived metrics.

In the present work, we calculate the RF, GWP and GTP using output from each individual model in the HTAP ensemble. We then compare our results to those obtained with the ensemble- mean method of Fry et al. (2012). Hence, we can quantitatively assess the extent to which the RF calculated with the mean fields accurately represents the mean of the RF calculated using the ozone fields from each model individually. Further, by comparing the estimates of model and metric uncertainty (as represented by the standard deviations) in RF, and in GWP and GTP, we can assess whether such a representative subset can be used to accurately convey the spread in derived climate metrics. The result of this assessment will then guide the extent to which the use of the computationally less expensive ensemble-mean fields can be used, without compromising the quality of information.

The particular case of NOx is interesting because cancellation between RF due to different components of the total RF (and hence the GWP and GTP) can substantially reduce model spread (Holmes et al., 2011), if individual components are correlated. Using values drawn from the aviation NOx literature, they found that in general, a large (positive) RF due to the short-lived ozone forcing (driven directly by the NOx) in any one model, was associated with a large (negative) long-lived ozone forcing (driven indirectly by the effect of NOx on methane concentrations) in the same model. Hence the uncertainty in the net RF, derived from considering the uncertainty in each component on its own, was found to be almost double the uncertainty in the net RF when the correlation was taken into account. Our work builds on Holmes et al. (2011) by exploiting results from a single multi-model intercomparison, and investigating the effects of different timescales on the cancellation, for emissions from a number of different regions, and extends it to CO and VOC (where the cancellation present in the NOx case is not present.

Section 2 introduces the HTAP data and scenarios, and describes the radiation code used to perform the radiative transfer calculations. The method of Fry et al. (2012) to generate the subset of fields for input to the radiation code is briefly described, together with a description of further preparing this output for generation of the GWP and GTP metrics. Section 3 presents the initial ozone and methane fields that serve as input to the radiation code for both methodologies, and briefly discusses their differences. Sections 4 and 5 discuss the effect of the different methodologies on the reported RF and GWP and GTP respectively, and conclusions are given in Section 6.

SR.2 (p27199 L1- p27202 L9)

2.1 Models

The HTAP study perturbation scenarios reduced by 20% emissions of short-lived ozone precursor gases NOx, CO and VOC in four different regions (North America, Europe, South Asia and East Asia), and a further run in which methane concentrations were perturbed globally. There are therefore 13 scenarios in addition to one control simulation. The models each ran for a period of 12 months after a spin-up time of at least 3 months (Fiore et al. (2009)). The resulting output of interest to this study are the tropospheric ozone fields, which are provided on each model grid at monthly mean resolution. Auxiliary information on methane lifetime changes for each scenario is used to calculate the change in methane and long-lived ozone concentrations as described in Section 2.3.

Table 1 shows the HTAP nomenclature for the experiments, and the locations of the source regions. 11 CTMs (see Table 2) produced results for these scenarios. For comparison with previous literature, the 11 models used in our study are the same as those used in Fry et al. (2012) and Collins et al. (2013) (Table 2).

Of the 11 CTMs used in this study, 9 use meteorological background fields from reanalyses to drive the model, while two (STOC-HadAM3-v01 and UM-CAM-v01) are coupled to global climate models (GCMs) and use 2001 sea ice and sea surface temperature data to drive the GCM. The models also use a variety of sources for the baseline emissions data, with the result that a 20% decrease in emissions is not equivalent in mass terms between models. Therefore, the model spread accounts for not only the uncertainties associated with transport and atmospheric chemistry, but also in background emissions, which can be a substantial source of uncertainty. As input to the radiation code, however, it is the absolute mass change of the species which is important for the radiative transfer calculations.

The model output is re-gridded to a common resolution of 2.75o latitude x 3.75o longitude, with 24 vertical levels, which is comparable to the resolution of the models on average. A common tropopause was identified as the level at which the lapse rate falls below 2 K km^{-1}. As many of the models do not include stratospheric chemistry, stratospheric changes in all species are neglected, and, above the tropopause, the models share a common climatology. Given the relatively coarse vertical resolution of the models, and that the data are monthly mean, any definition of tropopause is necessarily imperfect; however, this method ensures clarity when averaging monthly mean fields to form ensemble means, and minimises the aliasing of stratospheric ozone into the troposphere as part of the averaging process.

For each model, January, April, July and October are used as input to the code, in order to reduce run-time constraints whilst remaining sufficient to reasonably sample the annual cycle in transport and RF. Sensitivity tests have shown that the long-lived ozone and methane RFs are almost completely insensitive to increasing the number of months included (less than 1 part in 1000), and the short-lived ozone RFs have a sensitivity of the order of 0.5% to increasing the number of months. Table S4 provides a brief outline of the sensitivity tests.

2.2 Radiation code

This study uses the Edwards-Slingo radiation code (Edwards and Slingo, 1996). The code uses the two stream approximation to calculate radiative transfer through the atmosphere. Clouds are included in the code. Nine broadband channels in the longwave and 6 channels in the shortwave are used. Incoming solar radiation at mid-month, and Gaussian integration over 6 intervals is used to simulate variation in the diurnal cycle.

A common background climatology supplying temperature and humidity are taken from the European Centre for Medium-Range Weather Forecasts reanalyses (Dee et al., 2011). Mean cloud properties from the International Satellite Cloud Climatology Project (ISCCP) are also

used for all RF simulations (Rossow and Schiffer, 1999). RF is calculated as the difference in the net flux at the tropopause after the stratospheric temperature has been allowed to adjust using the standard fixed dynamical heating method (e.g. Fels et al. (1980)).

2.3 Construction of input metrics

The necessary inputs to the radiation code are the changes in atmospheric concentration of any radiatively active species. In this case, the relevant species are short-lived ozone, methane, and long-lived ozone, which is perturbed as a result of the influence of methane on the abundance of the OH radical.

The CTMs produce [OH],[O3] and associated atmospheric loss rates as 3-D output fields. Short-lived ozone can be used directly as input to the radiation code. Methane fields for each model and each simulation were globally homogeneous, and fixed at 1760 ppbv, except in the CH4 scenario, when they are reduced to 1408 ppbv. Equilibrium methane concentrations for each scenario have been calculated in Collins et al. (2013) from the change in methane life-time, Δα, as [CH4]=1760×(αcontrol+Δα)f, where the methane lifetimes are calculated in αcontrol Fiore et al. (2009). These lifetimes include loss terms such as those to soil processes; however all those except the atmospheric term are assumed to be constant. The change in methane life- time is also calculated in Collins et al. (2013) from the change in [OH] (since the atmospheric OH sink accounts for around 90% of loss of atmospheric CH4, and surface sinks are considered constant). Finally, the feedback factor, f is determined in Fiore et al. (2009) from the change in loss rates between the control and the CH4 perturbation scenarios, and accounts for the effect of methane change on its own lifetime (Prather, 1996).

Further, long-lived changes also arise from the change in ozone resulting from a change in methane, which in turn depends on the change in methane lifetime for a given scenario. The long-lived ozone changes for each model and scenario are calculated as described in West et al. (2009) by scaling the ozone change in the CH4 perturbation simulation by the relative change in methane concentration in each scenario as given in Fiore et al. (2009).

For each individual model, the inputs to the radiation code are the control and scenario 3- D monthly mean short-lived ozone, methane and long-lived ozone fields. Radiative transfer calculations are performed separately on each of these fields, so that the individual contributions can be separated out. The RF is the difference between the scenario and control fields for each species, and the total RF is taken to be the sum of these components. Sensitivity tests have shown that the total RF is very close (within 0.5%) to the sum of the individual contributions from the component gases. The mean of the resulting RF ensemble is denoted RF. This full model ensemble is contrasted with the method used in Fry et al. (2012). This method first constructs a representative subset of model input fields for input into the radiation code. This subset comprises the ensemble mean control fields, plus the ensemble mean ± standard deviation short-lived ozone, methane and long-lived ozone perturbations. This subset of fields is constructed as follows: Firstly, each model field for each month is regridded to a common resolution; secondly, the mean and standard deviation of the ozone field is calculated for each month, for each pixel at each level. The standard deviation is then added to or subtracted from the mean field to give a 3-D representative field for each month.

These fields are grouped into four cases; the first comprises the control fields; the second the mean total ozone change (i.e. the sum of the short- and long-lived mean ozone fields) together with the mean methane change; the third the mean plus standard deviation total ozone and methane change; and the final case the mean minus the standard deviation changes. Therefore the radiation code must run only three times for each HTAP scenario (plus once for the control run), relative to 33 (11 models, 3 gaseous species) plus 11 control runs for the complete case.

The subsetting method of calculation used in Fry et al. (2012) gives only the total RF for each scenario as output. The contributions to the total RF from each of the short-lived ozone, methane and long-lived ozone are then calculated from this total. First, the methane RF is calculated from the change in concentration using the simple formula of Myhre et al. (1998)

$$\Delta F -\alpha(\sqrt{M} -\sqrt{M0})-(f(M,N0)-f(M0,N0)) \quad (1)$$

where f(M,N) = 0.47ln[1+2.01×10−5(MN)0.75 +5.31×10−15M(MN)1.52], α is a constant, 0.12, N is N2O in ppb (constant at 315 ppb) and M is CH4 in ppb and the subscript 0 indicates the unperturbed case.

The difference between the total RF and this methane RF is then attributed to ozone. For the calculation of the GWP and GTP metrics, it is further necessary to separate the ozone RF between the short- and long-lived components. This is achieved by scaling the RF due to the (purely long-lived) ozone perturbations in the SR2 scenario by the ratio of the long-lived ozone change in any given scenario and the SR2 scenario. This RF is attributed to the long-lived ozone, with the final residual being attributed to the short-lived ozone. The mean and standard deviation of the RF calculated using this subset of fields are denoted RFEN.

2.4 Climate metrics

The methodology for calculation of the climate metrics (GWPs and GTPs) follows that described in Fuglestvedt et al. (2010), including the same impulse-response function for carbon dioxide, and the climate impulse-response function sensitivities from Boucher and Reddy (2008) which is needed for the GTP calculation. The metric calculations require the RF per unit emission per year, for each precursor and for the short-lived ozone, long-lived ozone and methane changes individually.

The calculation of GWP and GTP for each individual model is straightforward, as is the subsequent calculation of the ensemble mean and standard deviation. The implied change in methane emissions in the SR2 scenario must be calculated, as the scenario itself perturbed the atmospheric methane concentrations directly. This is done following the method in Collins et al. (2013) for each individual model.

For the Fry-method subset, the metrics must be constructed more carefully. We follow the method described in Collins et al. (2013). The GWP and GTP are both the sum of a short-lived ozone component, which depends only on the ozone RF, and a long-lived component, which depends on the methane and long-lived ozone RF, and the change in the methane lifetime. The ensemble-mean GWP and GTP are first calculated, and then a separate standard deviation due to each of the four variables is calculated. The total mean and standard deviation due to ozone changes are calculated, and then the total standard deviation is calculated in standard fashion as the square root of the sum of the variances. Note that this assumes independence between the variables. This is not necessarily the case because of correlations between the different perturbations (e.g Wild et al. (2001)); however for the purposes of this evaluation this provides a useful upper bound, and is consistent with the published literature (Collins et al. (2013)). The implied methane burden, which is necessary for normalising the RF in the SR2 scenario, is calculated from the methane lifetime and change in methane lifetime as described in Collins et al. (2013).

SR.3 (p27204 L22 – p27208 L27)

The major part of this section discusses the effect of the two averaging methods on the mean and spread of RF estimates. However, the RF's for the individual models in the HTAP ensemble have not previously presented, and may be of some interest. A brief discussion of the complete ensemble also serves to frame the subsequent discussion around appropriate averaging methods.

Figure 3 shows the RF for all 11 models, normalised by the change in burden of the emitted species (mW m−2(Tg year−1)−1 N, C, CO or CH4 for the SR3, SR4, SR5 and SR2 scenarios, respectively). RF due to short-lived ozone, methane and long-lived ozone is in general largest in SA and smallest in EU for any given model and scenario, largely due to an increased RF per unit radiatively active species due to warmer background temperatures in SA relative to EU, although non-linear chemical effects also affect the overall response (e.g. West et al. (2009)).

For VOC and CO, the methane and ozone RF act in the same direction, in contrast to NOx, where methane is suppressed and therefore it, and the long-lived ozone, act to oppose the RF due to short-lived ozone. The global-mean RF for any given model is less dependent on the location of the emission for the CO case than for the VOC or NOx case, as CO has a much

longer atmospheric residence time of 3 months, which is of the same order as the hemispheric atmospheric mixing time. The differences between the regions are therefore more pronounced for NOx than for VOC or CO, as a result of the greater inhomogeneity in the input fields.

The forcing for the CH4 perturbation scenario (bottom panel of Figure 3) comprises only the methane and long-lived ozone contributions, since there is no short-lived ozone forcing arising from a change in methane. The absolute methane RF is identical (-0.14 W m−2) across all models, as they all have the same mixing ratio change, but they differ in the size of the long-lived ozone response to the change in methane.

For a particular precursor species, models with a large response in one region will tend to have a large response in all regions, i.e. the models all agree on the order of the regional responses. These depend on the relative size of emissions change in each region and the mass-normalised RF. This is a good indicator of consistency across different emissions datasets and in transport in models, which information cannot be gained by using the model ensemble mean alone. For NOx, there is substantial correlation between the short-lived ozone and methane responses, and hence the short-lived and long-lived ozone responses, with r2 values between 0.70 (EA) and 0.86 (NA and SA, Table S2). This will result in a smaller standard deviation than if the quantities were truly independent of each other, as found by Holmes et al. (2011) for the case of aviation NOx emissions.

4.1 Ensemble-mean RF measures

Table 3 compares RF, ± 1 standard deviation per unit mass emissions change, with the mean and standard deviation of the computationally much less intensive RF_{EN} (the case in which the subsetting approach used in Fry et al. (2012) has been followed).

Differences between the means are only of the order of a few percent, with the largest differences found for the NOx NA case of 2%. For VOC and CO, the differences are essentially negligible. The larger fractional difference in the case of NOx is due to the fact that the means are a small residual of two much larger components. Hence RF_{EN} is representative of the true ensemble mean, RF. By contrast the standard deviation in the RF case is smaller for every scenario relative to RF_{EN}. This is largely associated with the inability of the pre-calculated ensemble mean fields to represent the true model spread, as described in Section 3.

Figure 4 separates the total RF into components due to the long-lived ozone, methane, and short-lived ozone contributions, for each scenario and gas, for the RF_{EN} and RF and their associated standard deviations. The differences in the size of the standard deviation is in general much larger for the short-lived ozone RF estimates (light blue bars), than for the long-lived ozone or methane components. This difference is, in effect a direct transform of the mathematical averaging effect in the input fields (see Section 2.3), and the standard error (i.e. the standard deviation divided by the mean) is the same in the input fields as it is after the radiative transfer calculations.

In the CH4 perturbation case, the absolute methane RFs (red bars) have no uncertainty associated with inter-model differences because the methane concentration change is fixed. The RF calculated using the formula of Myhre et al. (1998) is -139.6 mW m−2 for RF_{EN}, whereas the value calculated by the Edwards-Slingo radiation code for RF is slightly more negative at -141 mW m−2. The uncertainty bars arise from the variability in the implied methane burden change, which in turn arises from variability in the methane lifetime and change in methane lifetime.

SR.4 (p27209 L1 – p27212 L6)

5.1 Global warming potentials

The results above suggest that the subsetting approach to reduce the volume of calculations that must be performed may indeed be a useful method for quickly calculating ensemble mean RF; however, it is also apparent that estimates of the model spread might not be most appropriately calculated in this fashion. Metrics that are further downstream in terms of the impact chain, such as GWP and GTP, introduce further nonlinearities which must be considered when discussing the validity of this subsetting approach. Estimates of the GWP using the ensemble mean subsetting method are denoted GWPEN, while the the true values are denoted GWP.

GWPs for each individual model are calculated as described in Section 2.4 using the method of Fuglestvedt et al. (2010). Tables 4 and 5 give the values of the 20- and 100-year GWP respectively, in each case for the two methods under consideration, with the associated standard deviations. As previously, the mean values resulting from both methods remain very similar, with differences of the order of 2-3% for CO, 5% for VOC and up to 50% for NOx, once again as it is a small residual of the opposing short-and long-lived terms.

Estimates of the standard deviation using the subsetting method described in Fry et al. (2012), consistent with the previous section, are larger than the full model ensemble; however, the difference between the two standard deviation estimates is no longer simply related to the differences in the input fields. The total GWP at time horizon H is the sum of contributions from short- and long-lived components (i.e from RF due to short-lived ozone, and due to long-lived ozone, methane concentration and methane lifetime respectively). The difference between this estimate of the standard deviation and the full ensemble estimate therefore depends on the size of each of these terms and their relative contribution to the total estimate of the standard deviation. The absolute GWP of the short-lived ozone component does not depend on the time horizon under consideration, and it is still in effect directly proportional to the RF. Therefore the standard error (i.e. standard deviation/mean) of the short-lived ozone GWP remains the same as that for the RF and indeed for the input ozone fields, as does the relative difference in the size of the standard deviation estimates from the two methods. Table S3 gives the GWPs and GTPs, together with their associated standard deviation estimates for the total and for each contributing component.

The time-evolving components of the GWP, however, do not preserve this relationship, although the calculated standard deviations for each component remain larger using the subsetting method than calculating the true spread from the individual model GWPs. The total GWP is the sum of these components, and the relative difference in the calculated standard deviations from the two methods depends on the relative size of the contributions from the long- and short-lived components.

At 20 years, the short-lived ozone contributes proportionately more to the total GWP than at 100 years. This results in the relative differences between the standard deviation estimates from the two methods being proportionately larger at 20 than at 100 years for CO, VOC and NOx.

5.2 Global temperature-change potentials

The 20- and 100-year GTP means and standard deviations for the two methods are given in Tables 6 and 7. In common with the GWP calculations, the mean GTP’s for both methods differ by only a few percent. The standard deviation estimates resulting from the subsetting method are once again always larger than the true value obtained from the complete ensemble.

Similar principles apply to the relationship between the uncertainty estimates for the GTP as for the GWP. One important difference relative to the GWP in the 20-year case is the much larger relative contribution of the long-lived terms relative to the short-lived ozone terms. This means that, in contrast to the 20-year GWP, the 20-year NOx GTP is robustly negative in all cases.

For the 100-year GTP, the short-lived ozone contribution is a relatively larger contributor to the total than for the 20-year case. The relative contributions of each species and the methane lifetime to the total standard deviation estimates for both methods are given in the Supplementary material Table S3. This interplay between the various timescales associated with the GWP and GTP evolves with time, with the result that the difference between the two methods also evolves with time.