We would like to thank the 3 reviewers and the editor for their time in considering the paper and for their constructive comments. Our responses to these comments (repeated in blue) are below.

Response to reviewer 1

The manuscript is very well written and concise which makes it easy to follow the authors arguments. In addition, the authors openly address several potential shortcomings of their approach and included an extensive analysis of the performance based on cross-validation, which I find to be an excellent example of best practice.

Thank you for such a kind comment!

L1: I don't think that there is a single, agreed-on 'current method' for averaging model ensembles in climate science.

This is a fair point to make. We have changed the first sentence to highlight that standard averaging is commonly used, whilst not implying that it is the community standard.

Was:
The current method for averaging model ensembles, which is to calculate a multi model mean, assumes model independence and equal model skill.

Changed to:
Calculating a multi model mean, a commonly used method for ensemble averaging, assumes model independence and equal model skill.

L70: This is a minor point but I'm just pointing out that the REA does not down-weight models if they are more similar (dependent). It rather does the opposite and gives models which are closer to the multi-model mean additional weight as they are considered to be more reliable.

We appreciate the clarification. We have changed the sentence to reflect what you have said. It now reads:

Additionally, reliability ensemble averaging (REA) (Giorgi and Mearns, 2002) is an alternative weighting technique which instead gives higher weights to those models near the multi model mean.

L90: There are several more recent papers addressing and further developing this weighting method: http://doi.wiley.com/10.1029/2017JD027992 https://doi.org/10.1088/1748-9326/ab492f

Thank you for the heads up, these are now included.

Equation 1: The sum in the denominator should run over j!=i I assume?

Yes! Thank you for spotting that.

Section 2.1/Figure 1: I fail to understand how the sigma values where chosen from the figure and the text in this section. Can the authors elaborate on that please?

The purpose of Figure 1 is to illustrate how the weight depends on both the independence and the performance for a given value of sigma_d and sigma_s. These are not the sigma values used in the paper, it is instead a toy data set.

We deliberately didn't include too much description on the method of determining the sigma values, as we appreciate that there isn't a particularly objective way of doing so, as noted in by Knutti et al. (2017). In our case, to determine the values, we treated it somewhat like a machine learning problem, by having training and testing sets of data which don't overlap. The training set in this case was the refC1SD data including the total ozone column projection. The sigma values were found by optimising the weights, such that when they were applied to the refC1SD total ozone column they were a good fit to the observations. The testing set is then the weights applied to refC2 projections, which can be tested temporally out of sample (2010-2016), which avoids performing testing on data that has already been seen.

There may well be alternative values of sigma that deliver different fits to the observations and different scores in a perfect model test. But we believe that these sigma values balance performance and independence, and are found in a fair way.

The above information is reflected in a new short paragraph at the end of section 2.1 (from Line 126 in the new manuscript) that clarifies the choosing of the values of sigma.

L231: delete 'trend'? Figure 2 shows a time series of TCO not a trend right?

It is a trend in the sense that it is a smoothed time series with the inter-annual variability removed. The process for creating the trend is described in the paragraph beginning line 163. The term used by Scinocca et al. (2010) (the TSAM paper) is 'individual model trend' and for continuity we prefer to use that terminology.

L252: Figure 2→Figure 3

Indeed. Thanks!

L268: Maybe give the weight here relative to equal weighting similar to the lowest weight later in the paragraph. This seems to be the more meaningful metric as the absolute weight depends on the ensemble size.

Yes, we think that would add clarity, and possibly reassure readers that a weighting of 0.27 is not actually that large. The bold section has been added:

…CNRM-CM5-3, which has a weight of 0.27 **(297% the value of a uniform weighting.)**

Figure 4: This might just be an ambiguity in the language, so let me see if I understand correctly: Given a perfect model X, there is an unweighted projection Y and a weighted projection Y'

correct? So what is shown here is the difference abs(X-Y)-abs(X-Y')? If that is they case maybe rather call it the improvement of the (weighted) average than the average improvement? (except the last bar). I was a bit confused about the "Average improvement" which seems to indicate that this is showing some kind of average over different improvements (if my interpretation is correct).

The average here ('The average improvement in the Antarctic October TCO projection') is used because it is an average over the time series. You are correct, the improvement is the abs(multi model mean – pseudo truth) - abs(weighted mean – pseudo truth). We do see your point about ambiguity. To clarify that the average is temporal, we have changed 'average improvement' to '**mean monthly** improvement', in both the text and the y axis label of figure 4.

L296: 'alike any two models are' I don't think that this should be the aim or at least it should be worded more carefully. Two models can be 'alike' for different reasons, one of them being that they simulate the same system and are both 'good' at it so that the converge towards the truth. Or in other words: models should not be punished for arriving at the same answer independently. The assumption behind independence weighting schemes is that an inference on the model dependence can be made based on their output. If, e.g., two models share biases in several metrics this could be a sign that they also share one or several components.

Although, as you say, models shouldn't be punished for reaching the right answer independently, we should be more cautious of models which generate the right ozone column output but fail to simulate well any of the processes that are important in simulating ozone (e.g. temperature). That is one of the underlying reasons for analysing the models output this way: we want to have confidence that the models are simulating things for the right reason. This is important because we want to extend the weights onto the forecast scenario (refC2). By knowing that the models are simulating the ozone concentration well, because they simulate the underlying physics and chemistry well, we have more confidence that they will do so in the forecast.

Of course, if two very similar models both perform very well, we can see that they both simulate well for the right reason, because this will be reflected in the performance metrics. To some extent this can be seen in the temperature weighting of UMUKCA (Fig 3). UMUKCA has a pole significantly colder than the observations (and the other models) and resultantly gets heavily down weighted. It does not however receive a significant up weighting due to it being very different from the rest of the models.

Additionally, we have the ability to tune the sigma parameters. If we had a selection of near perfect models, the emphasis can be put on the performance aspect instead of the independence aspect.

L355: Have the authors tried to create the weights based on the refC2 simulations directly? Is it possible to receive sensible weights from that and how do they compare to the weights retrieved from the hindcasts?

Throughout the conception and refinement of this work we've had numerous discussions about what is the fairest set of simulations to create the weights from, and we hope we've highlighted the important parts of these discussions in the text.

To be specific, we have constructed weightings from both the refC1 simulations (essentially free running hindcasts) and refC2. In our opinion the weights from these were not sensible for a number of reasons:

- The refC1 set of simulations is not large and not many models ran multiple realisations. As a result, we don't have a good cover over the possible variable space, meaning that only a couple of models received a strong performance weighting. The same issue was true for the refC2 simulations. This free running non-smoothed set up also meant that the dependence weightings are particularly useful because of the large range of model output, for example the range of the total column ozone values can be as large as 250DU which is almost 100% of the 'normal' ozone column.
- To remove the interannual variability, to allow for a fair comparison, we would have to smooth metrics to allow for a trend. Missing data makes the smoothing tricky and means that we lose an amount of data.
- Essentially, we wanted to give the models the best chance of replicating the observations and the way to do this was to use the specified dynamics runs. This allows us to use all the available observational data and to test model response to events which would be lost through smoothing, such as volcanic eruptions and sudden stratospheric warming events. There is a brief discussion of this from L379 onwards.

Response to reviewer 2

Overall, this is well written concise paper and I think this is somewhat revised version of the manuscript.

Thank you. We very much appreciate your kind comments.

Major comment: Line 359 "The free running CCMI hindcast simulations (refC1) have a large...."

We've broken up your comments, so they can be addressed individually.

A) Which should be true for refC2 simulation as well, hence estimated ozone recovery dates should have much large uncertainty.

We have very carefully followed the work of Scinocca et al. (https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2009JD013622) when calculating the model trends and their associated uncertainties. This is a popular way for robustly extracting model trends whilst learning ensemble uncertainties. It is especially fitting for use here, given the methodology accounts for individual model weights. Having double checked the implementation of this method, we do believe that the uncertainties we calculate are genuine.

We clarify here that the confidence interval of the trend is for the trends with all interannual variability removed. The uncertainty presented by the confidence interval, is not a measure of the spread of the models in the ensemble. Instead it is a measure of the confidence in each of the individual model trends. For this reason, looking at the spread of the refC2 ensemble is not a good indicator of what the confidence interval of the multi model trend will be.

I think authors should give some clearer and better explanation for the selection of refC1SD over refC1 or first part of refC2 to calculate the weights. It is odd that weights are calculated for completely different dynamical space as ozone evolution would largely be determined by the changes in the stratospheric dynamics.

With reference to the metrics representing the relevant dynamical space, we believe the weights are generated from an appropriate set of metrics, which are relevant to ozone evolution the reasons for which are set out in section 3.2. With explicit reference to stratospheric dynamics, one of the metrics we include is based on the polar vortex which is extremely influential in southern polar dynamics in and around wintertime. We do highlight as well, that a better set of metrics could be chosen, but we are constrained by observational availability and model output.

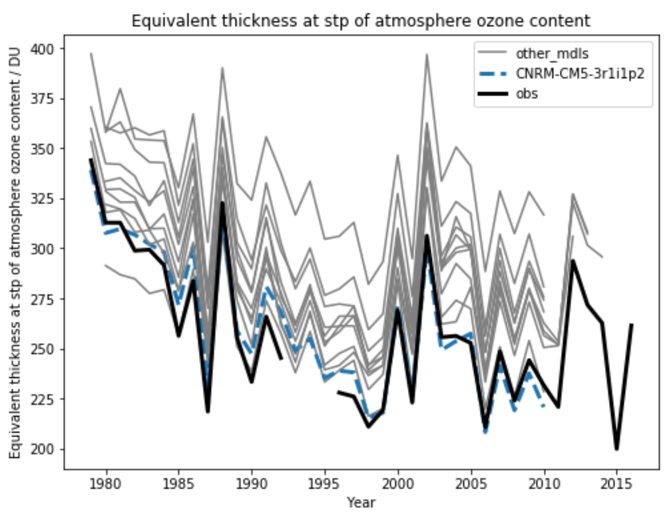
To justify the use of the weights generated using refC1SD on the refC2 simulations we performed both an out of sample test and a perfect model test (detailed in section 4.2). This shows that the weights learnt in one dynamical space are applicable to a different dynamical space. We also note that recent work from Orbe et al., (2020) that model biases that exist in a free running model can also exist in its specified dynamics counterpart (it is also true that it is partly down to the implementation of the specified dynamics). This would indicate that there is greater similarity between the dynamical spaces than you suggest, and therefore reason to use refC1SD for the selection weights.

Here, we repeat our response to a similar comment from the first reviewer, to further address the use of the refC1SD simulations. We did additionally construct weightings from both the refC1 simulations (essentially free running hindcasts) and refC2. In our opinion the weights from these would not be sensible for a number of reasons:

- The refC1 set of simulations is not large and not many models ran multiple realisations. As a result, we don’t have a good cover over the possible variable space, meaning that only a couple of models received a strong performance weighting. The same issue was true for the refC2 simulations. This free running non-smoothed set up also meant that the dependence weightings are particularly useful because of the large range of model output, for example the range of the total column ozone values can be as large as 250DU which is almost 100% of the ‘normal’ ozone column.
- To remove the interannual variability, to allow for a fair comparison, we would have to smooth metrics to allow for a trend. Missing data makes the smoothing tricky and means that we lose an amount of data.
- Essentially, we wanted to give the models the best chance of replicating the observations and the way to do this was to use the specified dynamics runs. This allows us to use all the available observational data and to test model response to events which would be lost through smoothing, such as volcanic eruptions and sudden stratospheric warming events.

For me higher weights to CNRM model over WACCM is really odd. Simple October TCO time series comparison (as well as ozone profile comparison) suggests CNRM being bit outlier. So, I am wondering high weightage to CNRM might be due to stronger nudging parameters. I think authors tried to explain in the paragraph starting at 353, but it is confusing and better explanation would help the readers.

Yes, CNRM does appear to be an outlier in the ensemble, but when we look at where the observations fit in the model spread (of CCMI refC1SD models) we see that the CNRM model is actually one of the closest to the observations (plot below). If the weights, as you suggest, are representative of higher nudging then we are up weighting the model which comes closest to the observations, which is a good thing. A way to test whether the nudging is having an effect on the weighting is to use the perfect model test.



In response to this comment, we have rewritten a large part of section 5 (L359-378) to clarify the utility of the perfect model test and why it shows that using nudged models doesn't have a detrimental effect on model weighting. The new text is below.

We generated weights from the refC1SD simulations which means that some metrics we chose are based on nudged variables, such as the lower stratospheric temperature gradient. As a result, one might expect that the model skill for these metrics should be equal, though given Fig. 3 this is not true. One may then expect that the weighting is not capturing model skill, but instead the skill of the models' nudging mechanisms; the models are nudged on different timescales ranging from 0.5 hr to 50 hr and from varying reanalysis products (Orbe et al., 2020). We use the perfect model test to show that the utility of the weighting methodology is not compromised by using models with such a variety in nudging time-scales and methods. As the perfect model test produces better projections, for models which are nudged in a variety of ways, we can conclude that the weighting is not dominated by nudging. Take for example UMUKCA-UCAM which is nudged quite differently compared to the ensemble, as evidenced by a southern pole significantly colder than the ensemble. When we take UMUKCA-UCAM as the pseudo truth (temporarily assuming the UMUKCA-UCAM output is the observational truth) we generate weights based

upon the refC1SD simulations and test them on the refC2 simulations. The weights generated are based on the dynamical system simulated in refC1SD which includes any model nudging. We can test how well these weights apply to a different dynamical system without nudging (refC2). The improvement in the WM compared to the MMM suggest that the weights generated from the refC1SD dynamical system do not predominantly reflect the quality of nudging and can be applied. If there hadn't been an improvement, then the dynamical systems described by refC1SD and refC2 may be too dissimilar for this weighting methodology and the weights may instead have been dominated by how well models are nudged. Nudging may be influencing the weights, but not to a degree that the accuracy of the projection suffers. Orbe et al. (2020) highlight the need for care when using the nudged simulations and we would like any future work on model weighting to quantify the impact of nudging upon model weights to reflect this.
We justified using the nudged refC1SD simulations, despite these considerations, for two reasons. Firstly, these nudged simulations give the models the best chance at matching the observational record, by providing relatively consistent meteorology across the models. The free running CCMI hindcast simulations (refC1) have a large ensemble variance and, despite producing potentially realistic atmospheric states, are not directly comparable to observational records. Secondly, the perfect model testing discussed above, demonstrates that the nudging doesn't have a detrimental effect on the model weighting.

B) Section 3.2: Please provide some more details about which pressure levels are used for lower stratospheric temperature, ozone.

- The ozone used for the projections (Fig 2) is used from the toz (total ozone) variable that the modelling groups output; we do not construct that.
- The temperature is taken as a weighted average over the lower stratosphere using the MSU TLS weighting function (i.e., a non-uniform average over a range of pressures). We have added a citation to the weighting function in section 3.2, in addition to the original citation in table 2. All further uses of temperature in any metric are constructed this way also. This is as described by Mears and Wentz (2009).
- The ozone used in the metrics is again total column ozone direct from the model output. The models do have different vertical ranges as discussed in section 3.1. We appreciate that we may have excluded some details readers were interested in. In section 3.2, any mention of 'ozone' has been changed to 'total column ozone'

Also, what does HCl averaged over whole stratosphere means? That does not make sense. Do you convert it in number density and calculate stratospheric column? Chlorine activation in the lower most stratosphere determines springtime ozone loss and mid-stratospheric or upper stratospheric HCl values are not that important.

Agreed, this is quite vague terminology. The HCl, as it is taken from the GOZCARDS product, has areas (both in vertical levels and latitude) that have missing data. So, we take all the model HCl output and regrid it to the same vertical and horizontal coordinates of the original HCl observations. These are constrained over a polar cap (90°S, 65°S) from levels between 316hPa and 15hPa (9 levels). For each spring season (SON) we take all the available data and create an average HCl concentration between 316hPa and 15hPa, from 90°S to 65°S. This is done in such a way that any data missing in the observations is excluded in the model data.

We have updated the hydrogen chloride part in Section 3.2 for clarity to add:

We consider a pressure range of 316 hPa to 15 hPa to capture the concentration in the lower stratosphere.

Minor comments:
i) Line 25: [e.g. Gillet, 2015, ..)

Good spot, thanks!

ii) Line 34: Ball et al., 2018 is not really good reference for that sentence.

Yes, fair point – reference removed.

iii) Table 2. Reference NIWA data V3.4 should be Bodeker et al., (2018) url={https://doi.org/10.5281/zenodo.1346424}

Thank you.

iv) line 327: that is not correct. In MMM, if model has more than one realization then generally individual model time series is created by calculating ensemble mean. If there is only one realization then most of the studies use 3 box-smoothing window.

True, though a naïve mean would average across all simulations. We have added the bold section below to reflect, say, HadGEM in CMIP5 having HadGEM2-AO, HadGEM2-CC and HadGEM2-ES:

Initially this may seem as if we are placing too much importance on one model, but consider that in a standard MMM, a model which runs three simulations **with different combinations of components** will have three times the influence of a model with a single simulation.

Response to reviewer 3

…but view the paper overall to be well-written, appropriate for the journal, and ready for publication after the authors have addressed the following minor comments.

Thank you!

L. 148 and throughout: What about when poor model performance manifests as poor simulation of the dynamics? If a model has an accurate chemical mechanism, maybe it looks good in the SD simulations, but it's poor when simulating the Antarctic vortex in free-running mode – doesn't that mean it should not be trusted to get the future ozone hole recovery right?

This is a very good question and raises a similar point to your next comment. Overall, in line with this comment and one from a different reviewer we have rewritten a large part of section 5

(L359-378) to clarify the utility of the perfect model test and why it shows that using nudged models doesn't have a detrimental effect on model weighting. A detailed response follows below.

Your question is essentially one about the nudging in the specified dynamic runs and whether this is influencing the measure of model performance. We have addressed this through the use of the perfect model test. The perfect model test pretends that the output from one model (in turn) is the "observational truth" (pseudo truth). This gives us the ability to have a full set of data both for the past (refC1SD) and the future (refC2). If what you're saying is true, that a model is only good in the past because it is nudged, then the perfect model test should show us this. For example, if we have a model which is strongly nudged in refC1SD so that it gets highly weighted because it is a good fit for the pseudo truth, when we test the fit of the weighted mean to the refC2 projections we will see that the weighted mean is actually a poor fit. In this case the perfect model test will show us that the weighted mean is not a good predictor of the future because the dynamical systems between refC1SD and refC2 are very different. However, if the weighted mean is a good predictor of the future (we measure compared to the multi-model mean), then we conclude that there is sufficient overlap (of the dynamical space) between the model in refC1 and refC2, and that the nudging hasn't had a large detrimental impact. The results of the perfect model test show that a weighted mean, constructed from refC1SD simulations, is a better projection for refC2 than a multi model mean and therefore, simulation performance is at least partly transitive between refC1SD and refC2. Additionally, we perform a small out of sample test with the overlap between observations and the beginning of the refC2 simulations (L287) which shows that the weights from the specified dynamics simulation, when applied to the free running simulation do improve the prediction. Care is taken here as it is only a small out of sample test, and so we use this result in addition to those from the perfect model test.

L. 194: How are the temperature metrics influenced by the SD versus free-running simulations? What would happen if you calculated the individual and total weights (i.e.,Fig. 3) using Ref-C2 instead of Ref-C1SD?

Here we address the use of the refC1SD simulations. We did additionally construct weightings from both the refC1 simulations (essentially free running hindcasts) and refC2 during the conception of this study. As per our response to other reviewers, in our opinion the weights from these were not sensible for a number of reasons:

- The refC1 set of simulations is not large and not many models ran multiple realisations. As a result, we don't have a good cover over the possible variable space, meaning that only a couple of models received a strong performance weighting. The same issue was true for the refC2 simulations. This free running non-smoothed set up also meant that the dependence weightings are particularly useful because of the large range of model output, for example the range of the total column ozone values can be as large as 250DU which is almost 100% of the 'normal' ozone column.
- To remove the interannual variability, to allow for a fair comparison, we would have to smooth metrics to allow for a trend. Missing data makes the smoothing tricky and means that we lose an amount of data.

- Essentially, we wanted to give the models the best chance of replicating the observations and the way to do this was to use the specified dynamics runs. This allows us to use all the available observational data and to test model response to events which would be lost through smoothing, such as volcanic eruptions and sudden stratospheric warming events.

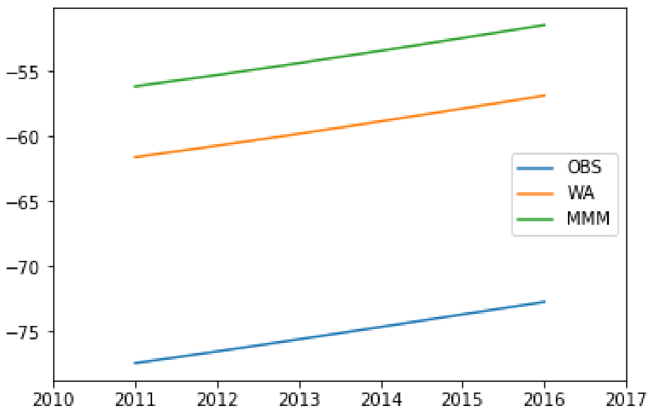
To justify the use of the weights generated using refC1SD on the refC2 simulations we performed both an out of sample test and a perfect model test (detailed in section 4.2). This shows that the weights learnt in one dynamical space are applicable to a different dynamical space.

And, similar to the previous comment, what if the nudging in the SD run is what causes a realistic decrease in temperature, not the coupling between decreased ozone and temperature (i.e., if ozone is poorly simulated, but the nudging imposes realistic temperature changes, will a high weighting be awarded to this model, for this metric, despite getting temperature “right for the wrong reasons”)?

As for your comment about L148, we believe that the perfect model test goes some way to checking that the nudging is not overly influencing the weighting. For example, take a model which has strong temperature nudging but in a free running scenario performs poorly. We would generate weights highly favouring this model over others in the ensemble, but these weights would create a poor projection. In the perfect model test this behaviour would be apparent, because the weighted projection will compare badly to the pseudo truth (a model which we have taken as the truth in order to test the methodology). However, the improvement in the WM compared to the MMM (shown by the perfect model tests) suggest that the weights generated from the refC1SD dynamical system do not predominantly reflect the quality of nudging and can be applied.

L. 290: Are the results of this out-of-sample test shown anywhere? It’s difficult for me to grasp what these RMSE values mean, in context, though I’d be curious to see the results of the test.

They have not, mainly because they don’t make for a particularly informative graph, and it would just be a zoomed in repeat of Figure 2. That figure does show the WM and MMM, alongside the observations, although admittedly not the RMSE values. The quick plot below shows the trends of the observations, the weighted mean (WA) and the multi model mean (MMM) for the out of sample period, where the y axis is the total ozone column (DU) relative to 1980 values.



We have given this thought, but don't think this plot (or a similar one for RMSE) adds much to the manuscript, and so don't include it.

L. 336: On the sensitivity of the final weightings to which performance metrics are included: You performed a dropout test, leaving one metric out at a time. But, what about leaving out two? For instance, since CNRM-CM5-3 stands out so significantly in its Total Weight, what if you leave out the two metrics where that model apparently excels above the other models in performance, i.e., Polar vortex breakdown trend and Ozone-temperature gradient? I am concerned that the Total Weight for each model is highly sensitive to the combination of observational metrics included, beyond what is tested by the single-metric dropout test.

This is a good question and we understand your concern, considering the uneven distribution of the weights for some metrics. However, performing some additional analysis, we don't find the uncertainty in recovery date to be particularly sensitive to this. If you leave out additional metrics you get a very similar uncertainty in recovery date:

- 95% confidence interval of recovery for 1 model 1 metric dropout: [2052.4, 2060.4]
- 95% confidence interval of recovery for 1 model 2 metrics dropout: [2052.1, 2060.7]
- 95% confidence interval of recovery for 1 model 3 metrics dropout: [2051.8, 2061.1]

So up to a point, the recovery date uncertainty is quite stable. However, if we progress further than 3 metrics being dropped, we begin to lose the point of why we're doing a multi metric weighting in the first place.

From this we conclude that the recovery date is not particularly sensitive to the inputs and believe this shows a level of robustness. We have added this into the manuscript at L259.

L. 50: "...and climate" should be "and climate forcers" or similar?

Changed

L. 200: "high" should be "highly" and "except of" should be "except for"

Changed, thanks.

L. 252: Should "Figure 2" instead be "Figure 3"?

Yes, thanks!

L. 267: “The total weighting, formed from the summation of individual metric weights...”Is this right? This conveys to me that all of the individual metric weights are simply totalled. Since CNRM has a couple weights >0.4, then its resulting total weight of 0.27 suggests this is not right...Found by Eq. 2 instead?

It is mentioned in section 2 that we “normalised over all the models to sum to 1”. For clarity we have changed summation to mean.

The total weighting, formed from the **mean** of individual metric weights **per model**, is largely influenced…

L. 273: “The lowest model weight is 55 % the value of a uniform weighting.” I understand this to mean that the lowest red bar in Fig. 3 is 55% the magnitude of the dashed black line, but the smallest bars (CCSRNIES-MIROC3.2 and EMAC-L47MA) look to be less than half, judging by eye. Is this statement correct?

Good spot, thanks! This was a typo (should be 45%), which has been changed. We have also double-checked other numeric values.

Projecting ozone hole recovery using an ensemble of chemistry-climate models weighted by model performance and independence

Matt Amos[1], Paul J. Young[1,2], J. Scott Hosking[3], Jean-François Lamarque[4], N. Luke Abraham[5,6], Hideharu Akiyoshi[7], Alexander T. Archibald[5,6], Slimane Bekki[8], Makoto Deushi[9], Patrick Jöckel[10], Douglas Kinnison[4], Ole Kirner[11], Markus Kunze[12], Marion Marchand[8], David A. Plummer[13], David Saint-Martin[14], Kengo Sudo[15, 16], Simone Tilmes[4], and Yousuke Yamashita[7]

[1]Lancaster University, Lancaster, UK

[2]Centre for Excellence in Environmental Data Science, Lancaster University, Lancaster, UK

[3]British Antarctic Survey, Cambridge, UK

[4]National Center for Atmospheric Research (NCAR), Boulder, Colorado, USA

[5]Department of Chemistry, University of Cambridge, Cambridge, UK

[6]National Centre for Atmospheric Science (NCAS), UK

[7]National Institute of Environmental Studies (NIES), Tsukuba, Japan

[8]LATMOS, Institut Pierre Simon Laplace (IPSL), Paris, France

[9]Meteorological Research Institute (MRI), Tsukuba, Japan

[10]Institut für Physik der Atmosphäre, Deutsches Zentrum für Luft- und Raumfahrt (DLR), Oberpfaffenhofen, Germany

[11]Steinbuch Centre for Computing, Karlsruhe Institute of Technology, Karlsruhe, Germany

[12]Institut für Meteorologie, Freie Universität Berlin, Berlin, Germany

[13]Environment and Climate Change Canada, Montréal, Canada

[14]CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

[15]Graduate School of Environmental Studies, Nagoya University, Nagoya, Japan

[16]Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokohama, Japan

Correspondence: Matt Amos (m.amos1@lancaster.ac.uk)

Abstract. ~~The current method for averaging model ensembles, which is to calculate~~ Calculating a multi model mean, a commonly used method for ensemble averaging, assumes model independence and equal model skill. Sharing of model components amongst families of models and research centres, conflated by growing ensemble size, means model independence cannot be assumed and is hard to quantify. We present a methodology to produce a weighted model ensemble projection, accounting for model performance and model independence. Model weights are calculated by comparing model hindcasts to a selection of metrics chosen for their physical relevance to the process or phenomena of interest. This weighting methodology is applied to the Chemistry-Climate Model Initiative (CCMI) ensemble, to investigate Antarctic ozone depletion and subsequent recovery. The weighted mean projects an ozone recovery to 1980 levels, by 2056 with a 95 % confidence interval (2052–2060), 4 years earlier than the most recent study. Perfect model testing and out-of-sample testing validate the results and show a greater projective skill than a standard multi model mean. Interestingly, the construction of a weighted mean also provides insight into model performance and dependence between the models. This weighting methodology is robust to both model and metric choices and therefore has potential applications throughout the climate and chemistry-climate modelling communities.

1 Introduction

Global chemistry-climate models (CCMs) are the most comprehensive tools to investigate how the global composition of the atmosphere develops, both naturally and under anthropogenic influence (Flato et al., 2014; Morgenstern et al., 2017; Young et al., 2018). As with projecting climate change, consensus views of the past and potential future evolution of atmospheric composition are obtained from coordinated CCM experiments (Eyring et al., 2008; Lamarque et al., 2013; Morgenstern et al., 2017) and subsequent analysis of the ensemble of simulations (Iglesias-Suarez et al., 2016; Dhomse et al., 2018). Although not a complete sample of structural and epistemic uncertainty, these ensembles are an important part of exploring and quantifying drivers of past and future change, and evaluating the success of policy interventions, such as stratospheric ozone recovery resulting from the Montreal Protocol and its amendments (Dhomse et al., 2018; WMO, 2018). Typically, analysis of an ensemble investigates the behaviour and characteristics of the multi-model mean and the inter-model variance (Solomon et al., 2007; Tebaldi and Knutti, 2007; Butchart et al., 2010), rather than accounting for individual model performance or lack of model independence (Knutti, 2010; Räisänen et al., 2010). Methods to address these shortcomings have been proposed for simulations of the physical climate ~~(Gillett, 2015; Knutti et al., 2017; Abramowitz et al., 2019, e.g.)~~(e.g., Gillett, 2015; Knutti et al., 2017; Abramowitz et al., 2019), but this topic has received less attention in the atmospheric composition community. Here, we demonstrate a weighting method for the CCM simulation of Antarctic ozone loss and projected recovery, where the weighting accounts for model skill and independence over specified metrics relevant to polar stratospheric ozone. We apply this to the recent Chemistry-Climate model initiative (CCMI) (Morgenstern et al., 2017) ensemble and demonstrate the impact of the weighting on estimated ozone hole recovery dates.

Many years of scientific studies and assessments have tied stratospheric ozone depletion to the anthropogenic emission and subsequent photochemistry of halogen-containing gases, such as chlorofluorcarbons (CFCs), hydrofluorocarbons (HCFCs) and halons (WMO, 2018). This science guided the development of the Montreal Protocol, and its subsequent amendments, to limit and ban the production of these ozone-destroying gases, and stratospheric ozone is now thought to be recovering ~~(Solomon et al., 2016; Chipperfield et al., 2017; Ball et al., 2018)~~(Solomon et al., 2016; Chipperfield et al., 2017). Of particular concern is the Antarctic "ozone hole": a steep decline in high latitude stratospheric ozone during austral spring that can reduce ozone concentrations to near zero at particular altitudes, driven by polar night-time chemistry, cold temperatures and heterogeneous catalysis on polar stratospheric clouds (PSCs) (Solomon, 1999). While the ozone hole continues to appear in each austral spring, it appears to be showing signs of recovery (Langematz et al., 2018). The strong cooling associated with Antarctic ozone depletion (Thompson and Solomon, 2002; Young et al., 2012) has driven circulation changes in the stratosphere and in the troposphere, particularly in austral summer. This has notably included an acceleration and poleward movement of the southern high latitude westerly winds and associated storm tracks (Son et al., 2008; Perlwitz et al., 2008), leading to summertime surface climate changes through many lower latitude regions including the tropics (Thompson et al., 2011).

The recovery process is slow due to the long atmospheric lifetimes of ozone depleting substances, and could be hampered by releases of ozone depleting substances (ODSs) not controlled by the Montreal Protocol, such as short-lived halogens (Claxton et al., 2019; Hossaini et al., 2019) or nitrous oxide (Portmann et al., 2012; Butler et al., 2016), or instances of non-compliance, such as the recent fugitive emissions of CFC-11 (Montzka et al., 2018; Rigby et al., 2019). Recovery itself is often defined as the date at which the ozone layer returns to its 1980 levels, and this is the benchmark used by the WMO (WMO, 2018) to assess the progress due to the implementation of the Montreal Protocol.

The assessment of when the ozone layer will recover is conducted using an ensemble of chemistry-climate models, forced by past and projected future emissions of ozone depleting substances (ODSs) and climate forcers(Eyring et al., 2010; Dhomse et al., 2018). Such ensembles are used to establish the robustness of the model results for a particular scenario: when several models agree, the prevailing assumption is that we can have greater confidence in the model projections. Yet, there has been much discussion about how true this assumption is (Tebaldi and Knutti, 2007; Sanderson et al., 2015b; Abramowitz et al., 2019). In an ideal scenario, every model within an ensemble would be independent and have some random error. In this case, we would expect that increasing the ensemble size would decrease the ensemble uncertainty and allow us to better constrain the mean value. However, in modern model inter-comparison projects this is not the case: although often developed independently, models are not truly independent, often sharing components and parametrisations (Knutti et al., 2013); models are not equally good at simulating the atmosphere (Reichler and Kim, 2008; Bellenger et al., 2014); and lastly, models do not have a predictable random error but instead have layers of uncertainty extending from uncertainties in parametrising sub-grid processes (Rybka and Tost, 2014) to structural uncertainties from the design of the model (Tebaldi and Knutti, 2007; Knutti, 2010).

Given these issues, there is currently no consensus on how best to combine model output when analysing an ensemble. Probably the most widely used and simplest is to take a multi model mean where each model contributes equally, and indeed it has also been established that an ensemble mean performs better than any single model (Gleckler et al., 2008; Reichler and Kim, 2008; Pincus et al., 2008; Knutti et al., 2010). A more sophisticated method is to weight individual ensemble members, accounting for model performance as well as the degree of a model's independence. Weighting methods of various forms have been developed and implemented on global physical climate model ensembles (Tebaldi et al., 2005; Räisänen et al., 2010;

Haughton et al., 2015; Knutti et al., 2017), but seldom for atmospheric composition. In most cases the weights are calculated from comparison of model hindcasts to observational data, either for a single variable of interest or over a suite of diagnostics. ~~In addition to these weighting techniques there are other methods for generating ensemble means, such as clustering (Yuan and Wood, 2012; Hyde et al., 2018) and~~ Additionally, reliability ensemble averaging (REA) (Giorgi and Mearns, 2002) is an alternative weighting technique which gives higher weights to those models near the multi model mean. The main motivation for using a weighted mean is to encapsulate model skill and model independence, such that we down-weight models which perform less well and/or are more similar.

Quantifying model skill (or performance) against comparable observations forms an important part of the validation and analysis of multi-model ensembles (Gleckler et al., 2008; Flato et al., 2014; Harrison et al., 2015; Hourdin et al., 2017; Young et al., 2018). Many CCM inter-comparison projects feature validation and assessment through the use of observation-based performance metrics, which may capture model performance for particular atmospheric variables (e.g., temperature, chemical species concentrations, jet position), or be a more derived quantity which gets closer to evaluating the model against the process it is trying to simulate (e.g., ozone trends vs. temperature trends, chemical species correlations, chemistry-meteorology/transport relationships) (Eyring et al., 2006; Waugh and Eyring, 2008; Christensen et al., 2010; Lee et al., 2015). Performance metrics are chosen based upon expert knowledge of the modelled system to ensure that metrics are highly related to the physical or chemical processes that the models are being evaluated on.

In this study we develop a weighting methodology, originally presented by Sanderson et al. (2017) and Knutti et al. (2017), for CCM ensembles that accounts for model performance and model independence. We apply it to the important issue of estimating Antarctic ozone recovery using several well-established metrics of model performances, where previously only unweighted means have been used. We first describe our weighting framework in Sect. 2, before describing the model and observational data in Sect. 3. Section 4 presents the application of the weighting framework to Antarctic ozone depletion and the corresponding results. Sections 5 and 6 present a summary and our conclusions.

2 The model weighting framework

In this study, we develop and exploit a framework to calculate model weights based on ~~the~~ recent work in the physical climate science community ~~(Sanderson et al., 2015a, b, 2017; Knutti et al., 2017)~~(Sanderson et al., 2015a, b, 2017; Knutti et al., 2017; Lorenz et a
Here, for an ensemble of N models, the weight for model i (w_i) is given by

$$w_i = \exp\left(-\frac{D_i^2}{n_i\sigma_D^2}\right) \Big/ \left(1+\sum\nolimits_{\cancel{i\neq j}\, j\neq i}^{N} \exp\left(-\frac{S_{ij}^2}{n_i\sigma_S^2}\right)\right). \quad (1)$$

The numerator captures the closeness of the model to observations. D_i^2 is the squared difference between a model and the corresponding observation, which is a measure of performance. The denominator captures the closeness of a model to all other models by comparing the squared difference between them (S_{ij}^2). Both σ_D and σ_S are constants which allow tuning of the weighting to preference either independence or performance (see discussion below). Put more simply, a model has a larger weighting if it closely matches observations and is suitably different to the other models in the ensemble. Finally, Eq. (1) differs from similar versions (e.g., Knutti et al., 2017) through the addition of n_i, which is the size of the data used to create the weighting. This could be the amount of grid points for a spatial field, the number of points in a time series, or just one for a single-valued statistic, and it normalises the data by length allowing for comparison between models and variables with time series of different length and time invariant parameters.

Investigating and evaluating a phenomenon or complex process often relies on identifying multiple metrics since it can only be partially expressed by any single variable. Expert understanding of the physical process is needed to select a set of relevant metrics with which to develop the process-based weighting. Including multiple metrics, provided they are not highly correlated, has the further benefit of giving less weight to models which perform well but do so for the wrong reasons. In this framework, ensuring that these metrics influence the weighting proportionally is done by normalising the model data using a min-max scaling between 0 and 1.

When combining multiple metrics into a weighting, the weight of the i^{th} model can be found from

$$w_i = \left(\sum_{k=1}^{M} \exp\left(-\frac{D_{ik}^2}{n_{ik}\sigma_D^2} \right) \right) \Big/ \left(M + \sum_{k=1}^{M} \sum_{j \neq i}^{N} \exp\left(-\frac{S_{ijk}^2}{n_{ik}\sigma_S^2} \right) \right), \tag{2}$$

where M is the total number of metrics and k is the index of the metric. Note that the summation is performed separately over the numerator and the denominator. This means that we calculate the performance and independence scores over all the metrics combined before merging the scores to create the final weighting which, as before, is normalised over all the models to sum to 1.

We take the combined weights for each model and apply them to our parameter or process of interest (the evolution of stratospheric ozone here). As with the metrics this parameter needn't be a time series and could be a spatial distribution or a single measure. The weighted projection is therefore $x = \sum_{i=1}^{N} w_i x_i$, where x_i is an individual model projection and w_i is the associated weight.

2.1 Choosing sigma values

The two scaling parameters (σ_S, σ_D) represent a length scale over which two models, or a model and observation, are deemed to be in good agreement. For example, a large σ_S would spread weight over a greater number of models as more models would lie within the length scale of σ_S. On the other hand, a small σ_S sets a higher tolerance for measuring similarity. The choice of the sigma values needs to be considered carefully to strike a balance between weighting all models equally, thus returning to a multi model mean, versus weighting just a few selected models. As the same values of sigma apply across all metrics it is necessary for the data to be normalised to the same values, ensuring that metrics impact the weightings equally. Figure 1 shows how the weighting function depends on σ_S, σ_D, model performance and model independence.

As noted in Knutti et al. (2017) there is not an objective way of determining optimal sigma values. Our method of selecting appropriate parameter values was to consider a training and a testing set of data, much like a machine learning problem. We determined the values of sigma using the training data, which in this case is the refC1SD simulations, such that the weighted training data gave a good fit to the observations. The testing data (refC2 simulations) allowed us to test the weights and sigma values out of the temporal range of the training data, which avoids performing testing on data that was used to tune the parameters.

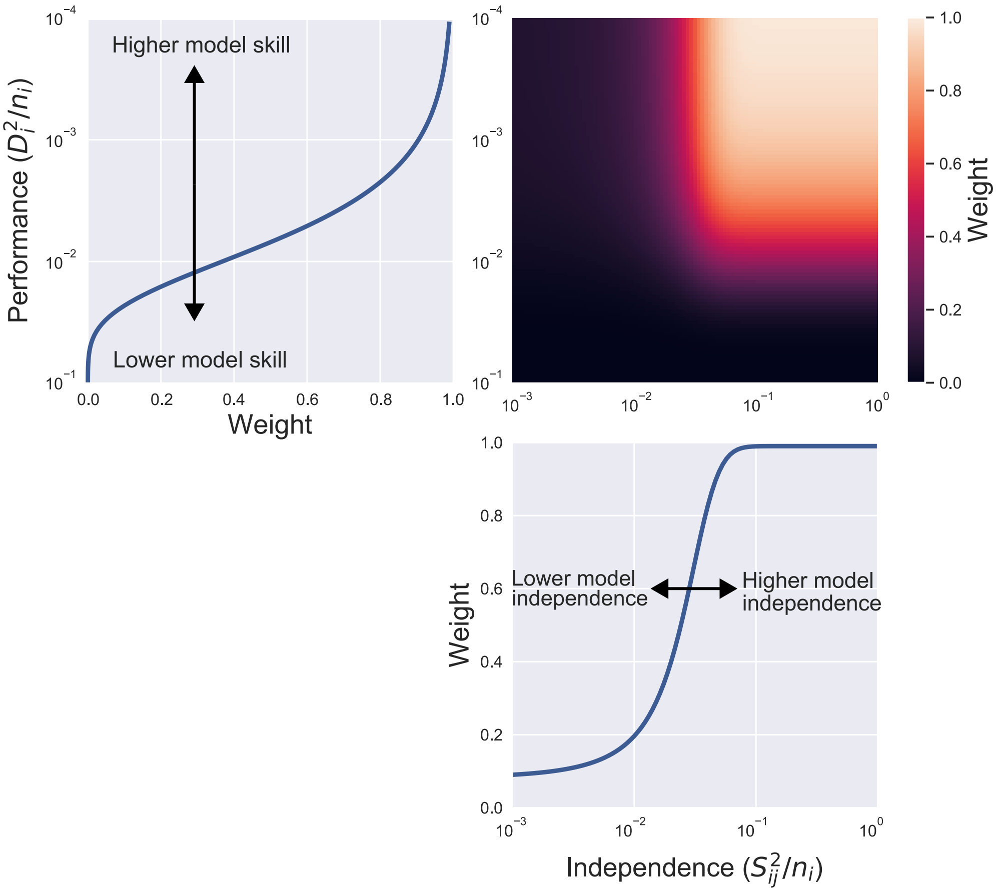


Figure 1. Top right shows the overall weighting function w_i (Eq. 1), plotted for 11 models ($N = 11$) with $\sigma_D = 0.1$ and $\sigma_S = 0.1$. Top left shows the contribution to the weighting due to model performance (at $S_{ij}^2/n_i = 1$) and bottom right show the contribution due to model independence (at $D_i^2/n_i = 10^{-4}$). A model which has higher independence and higher skill receives a larger weight. For the weight due to performance (top left) we can see that the weight equals e^{-1} when $D_i^2/n_i = \sigma_D^2$. This shows how σ_D acts as a length scale that determines how close a model has to be to observations to receive weight. σ_S works similarly, setting the length scale that determines similarity.

3 Applying the weighting framework to the Antarctic ozone hole

We demonstrate the applicability of this weighting framework by applying it to the important and well-understood phenomenon of the Antarctic stratospheric ‘ozone hole’, for which we can use several decades of suitable observations to weight the models.

Table 1. The CCMI model simulations used in this analysis and their key references.

Model	refC1SD realisation(s)	refC2 realisation(s)	Reference(s)
CCSRNIES-MIROC3.2	r1i1p1	r1i1p1	Imai et al. (2013), Akiyoshi et al. (2016)
CESM1-CAM4Chem	r1i1p1	r1i1p1	Tilmes et al. (2015)
CESM1-WACCM	r1i1p1	r1i1p1	Marsh et al. (2013), Solomon et al. (2015), Garcia et al. (2017)
CHASER-MIROC-ESM	r1i1p1	r1i1p1	Sudo et al. (2002), (Sudo and Akimoto, 2007), Watanabe et al. (2011), Sekiya and Sudo (2012) Sekiya and Sudo (2014)
CMAM	r1i1p1	r1i1p1	Jonsson et al. (2004), Scinocca et al. (2008)
CNRM-CM5-3	r1i1p2 r2i1p2[a]	r1i1p1	Michou et al. (2011), Voldoire et al. (2013)
EMAC-L47MA	r1i1p1 r1i1p2[a]	r1i1p1	Jöckel et al. (2010), Jöckel et al. (2016)
EMAC-L90MA	r1i1p1 r1i1p2[a]	r1i1p1	
IPSL	r1i1p1	r1i1p1	Marchand et al. (2012), Szopa et al. (2013), Dufresne et al. (2013)
MRI-ESM1r1	r1i1p1	r1i1p1	Deushi and Shibata (2011), Yukimoto (2011), Yukimoto et al. (2012)
UMUKCA-UCAM	r1i1p1	r1i1p1	Morgenstern et al. (2009), Bednarz et al. (2016)

[a] Represents the simulations used in the similarity analysis, but that did not form part of the model weighting.

Below, we describe the model and observation data used and the metrics selected, against which we measure model performance and independence.

3.1 Model and observation data sources

CCM output was taken from the simulations conducted under Phase 1 of the Chemistry-Climate Model Initiative (CCMI) ((Morgenstern et al., 2017) and refs. therein), which represents an ensemble of 20 state-of-the-art CCMs (where chemistry and atmospheric dynamics are coupled) and chemistry transport models (CTMs, where the dynamics drives the chemistry, but there is no coupling). A detailed description of the participating models is provided by Morgenstern et al. (2017), and here we briefly review their overarching features. Most models feature explicit tropospheric chemistry and have a similar complexity of stratospheric chemistry though there is some variation in the range of halogen source gases modelled. Horizontal resolution of the CCMs ranges from between $1.125°\times1.125°$to $5.6°\times5.6°$. Vertically, the atmosphere is simulated from the surface to near the stratopause by all models, and many also resolve higher in the atmosphere. Vertical resolution varies throughout the models, both in the number of levels (34 to 126) and their distribution. All models simulate the stratosphere, although they differ in whether they have been developed with a tropospheric or stratospheric science focus.

We focus on two sets of simulations, called refC1SD and refC2, and for the weighting analysis we only consider models which ran both simulations. Table 1 details the exact model simulations used. The refC1SD simulations cover 1980–2010 and represent the specified dynamics hindcast, where the models' meteorological fields are nudged to reanalysis datasets in order

Table 2. The observational products and respective variables used to construct metrics on which to weight the models.

Product	Variable	Metric/s	Citation
MSU	Lower stratosphere temperature (TLS)	TLS/TLS Gradient/Ozone-temperature	Mears and Wentz (2009)
NIWA-BS	Total Column Ozone V3.4 (TCO)	TCO gradient/Ozone-temperature	~~Bodeker et al. (2005)~~ Bodeker et
GOZCARDS	Hydrogen chloride concentration	Antarctic hydrogen chloride concentration	Froidevaux et al. (2015)
ERA-Interim	Eastward wind speed	Polar vortex breakdown trend	Berrisford et al. (2011)

~~The observational products and respective variables used to construct metrics on which to weight the models.~~

that the composition evolves more in line with the observed inter-annual variability of the atmosphere. In addition to being nudged by meteorology the refC1SD runs are forced by realistically varying boundary conditions, including greenhouse gas (GHG) concentrations, ODS emissions, and sea surface temperatures (SSTs) and sea-ice concentrations (SICs). The refC1SD simulations are used to create the model weightings since these are the models' best attempt at replicating the past, giving reasonable confidence that any down-weighting arises due to poorer model performance or strong inter-model similarity. It must be noted that the nudging process is not consistent across the models (Orbe et al., 2018) and we should be mindful that it has the capability to influence the weighting. We discuss the choice to use refC1SD simulations in greater detail in Sect. 5.

The refC2 simulations cover 1960–2100 and are used to construct weighted projections of Antarctic ozone recovery, using the weights calculated from refC1SD. The forcing from GHGs and anthropogenic emissions follows the historical scenario conditions prescribed for the fifth coupled model Inter-comparison project (CMIP5) (Lamarque et al., 2010) up to the year 2000, and subsequently follows representative concentration pathway (RCP) 6.0 for GHGs and tropospheric pollutant emissions (van Vuuren et al., 2011); the ODS emissions follow the World Meteorological Organisation (WMO) A1 halogen scenario (WMO, 2011). From CCMI this is the only scenario which estimates the future climate change and developments to stratospheric ozone.

Model performance was evaluated against a series of well-accepted metrics (see below), drawing from widely used observational and reanalysis datasets listed in Table 2. Assessing models and ensembles using observational data is a principal way of validating models (Eyring et al., 2006; Waugh and Eyring, 2008; Dhomse et al., 2018) and this is the methodology we follow, with the addition that we create the weights based upon this skill, alongside model independence.

Like many ozone recovery studies, we utilise TSAM (time series additive modelling) (Scinocca et al., 2010) to quantify projection confidence, which produces smooth estimates of the ozone trend whilst extracting information about the inter-annual variability. Here, the TSAM procedure involves finding individual model trends for the refC2 simulations by removing the inter-annual variability using a generalised additive model. Each model trend is then normalised to its own 1980 value. The weighted mean (WM) is created by summing model weights with individual model trends. Two uncertainty intervals are created: a 95 % confidence interval, where there is a 95 % chance that the WM lies within; and a 95 % prediction interval, which captures the uncertainty of the WM and the inter-annual variability.

3.2 Metric choices - How best to capture ozone depletion

The first step in the weighting process is to identify the most relevant processes that affect Antarctic ozone depletion to allow for appropriate metric choice. Suitable metrics require adequate observational coverage and for the models to have outputted the corresponding variables. The metrics we chose are as follows:

Total ozone column gradient. This is the first derivative with respect to time of the total ozone column. Given the discontinuity in the total ozone column record, the years 1992–1996 are excluded. It is a southern polar cap (60°S–90°S) average over austral spring (October and November). September is not included due to discontinuous coverage in the observations.

Lower stratosphere temperature. The lower stratosphere temperature for all of the models are constructed using the MSU TLS-weighting function (Mears and Wentz, 2009). The MSU dataset extends to 82.5°S, and therefore the southern polar cap average ranges from 60°S to 82.5°S and is temporally averaged over austral spring (Sept, Oct, Nov).

Lower stratosphere temperature gradient. This is the first derivative with respect to time of the lower stratospheric temperature found above.

Breakdown of the polar vortex. The vortex breakdown date is calculated as when the zonal mean wind at 60°S and 20 hPa transitions from eastward to westward as per Waugh and Eyring (2008). We find the trend of the breakdown date between the years 1980–2010 and the gradient of the trend forms the polar vortex breakdown metric.

Ozone-temperature gradient. Both the lower stratosphere temperature and the ~~ozone~~ total ozone column are separately averaged over 60°S to 82.5°S and the October and November mean was taken. We determined a linear relationship between temperature and ~~ozone~~ total ozone column and the gradient of this linear relationship forms the ozone-temperature metric (Young et al., 2013).

Ozone trend-temperature trend gradient. This is similar to the metric above except that we first calculated the time derivative of the ~~ozone~~ total ozone column and temperature polar time series before calculating the linear relationship. The gradient of the linear relationship is the ~~ozone~~ total ozone column trend temperature trend gradient metric.

Hydrogen chloride. The hydrogen chloride concentration was averaged over the austral spring months ~~, throughout the stratosphere,~~ and over the Southern Polar cap, for areas which have observational coverage. We consider a pressure range of 316 hPa to 15 hPa to capture the concentration in the lower stratosphere.

These metrics capture two of the main features of ozone depletion, namely: 1) the decrease in temperature over the poles caused by the depletion of ozone, and 2) the breakdown of the vortex which has a major role of isolating the ozone depleted air mass. The chlorine metric encapsulates the anthropogenic release of ODSs and the main chemical driver of ozone depletion. Ozone-temperature metrics allow us to look at model success in reproducing the temperature dependency in ozone reaction rates and stratospheric structure. By looking at the instantaneous rate of change as well as the overall trends, we can gather a picture of both short-term and long-term changes for a range of chemical and dynamical processes.

The metrics are not ~~high~~ highly correlated, except ~~of~~ for the total ozone column gradient and the lower stratosphere temperature gradient, which are correlated because of the strong coupling of ozone and temperature in the stratosphere (e.g., Thompson and Solomon, 2008). Although this could be cause to discard one of the metrics, to avoid potential double counting, we retain and use both to weight because the models may not necessarily demonstrate this coupling that we see in observations. By considering this variety of metrics, the approach aims to demonstrate that models do not just get the 'right' output, but that they do so for the right reasons.

3.3 Evaluating the weighting framework

Two types of testing were used to investigate the usefulness of the weighted prediction and to validate metric choices. Firstly, we performed a simple out-of-sample test on the weighted prediction against the total ozone column observations from NIWA-BS. Although the weights are generated from comparison between the specified dynamics runs (refC1SD) and observations, it does not necessarily follow that the weighted projection created using the free running (refC2) runs will be a good fit for the observations. To test this, we compared the refC2 multi model mean and weighted projection to the observations. Due to the large inter-annual variability in the total column ozone (TCO) observations, we do not expect the weighted average to be a perfect match; after all, free running models are not designed to replicate the past. However, we need to test the level of agreement between the weighted mean and the observations for an out-of-sample period (2010–2016). This serves a secondary purpose of determining transitivity between the two model scenarios used: i.e., that the weightings found from refC1SD apply to refC2.

Secondly, we used a perfect model test (also known as model-as-truth or a pseudo model test) to determine whether our weighting methodology is producing valid and robust projections. In turn, each model is taken as the pseudo truth and weightings are found in the same way as described in Sect. 2 except the pseudo truth is used in place of observations. From these weightings we can examine the skill with which the weighted mean compares to the pseudo truth. We are normally limited to a single suite of observations, but a perfect model test allows us to test our methodology numerous times using different pseudo truths, demonstrating robustness.

Perfect model testing also allows us to test transitivity between scenarios since, unlike with the obvious temporal limit on observations, the pseudo truth exists in both the hindcast and forecast. If a weighting strategy produces weighted means which are closer to the pseudo truth than a multi model mean, then we can have some confidence that we can apply a weighting across model scenarios. Herger et al. (2019) compare the perfect model test to the cross validation employed in statistics, but note that although necessary, perfect model tests are not sufficient to fully show out-of-sample skill which in this case is scenario transitivity. It should be backed up by out-of-sample testing as described above.

4 Applying the weighting framework to Antarctic ozone simulations

4.1 Antarctic ozone and recovery dates

Figure 2 shows the October weighted mean (WM) total column ozone (TCO) trend from the refC2 simulations for the Antarctic (60–90°S). The weights are calculated using Eq. (2), and are based on both model performance and independence. All models

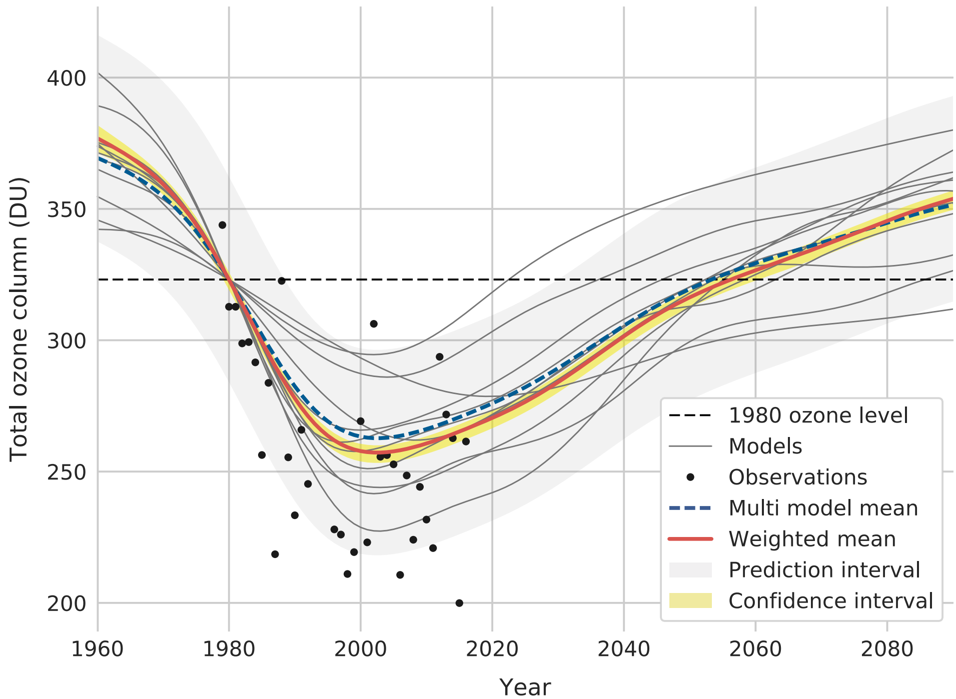


Figure 2. Antarctic (60–90°S) October TCO. The weighted mean (refC2 simulations weighted upon refC1SD performance and independence) is shown in red, the multi model mean (refC2 simulations) is shown in blue, and individual refC2 model trends are shown in grey. The NIWA-BS observations are shown in black. All model projections and ensemble projections are normalised to the observational 1979–1981 mean shown as the black dashed line. 95 % confidence and prediction intervals for the weighted mean are also shown with shading.

simulate ozone depletion and subsequent recovery but with large discrepancies in the absolute TCO values and the expected recovery to 1980 levels (see Dhomse et al., 2018), from here on referred to as D18. The WM and multi model mean (MMM) are similar, given the small number of models considered from the ensemble ($N = 11$). At maximum ozone depletion, around the year 2000, the WM projects a significantly lower ozone concentration (5 DU) than the MMM. This steeper ozone depletion seen in the WM fits the observations better than the MMM, although the modelled inter-annual variability seems to under predict the observations.

The WM predicts a return to 1980 TCO levels by 2056 with a 95 % confidence interval (2052–2060). For comparison the recovery dates presented in D18 were 2062 with a 1σ spread of (2051–2082). Although taken from the same model ensemble (CCMI), the subset of models in this analysis is smaller than that used in D18 meaning that difference in recovery dates between the two works is attributable to both the methodology and the models considered. The smaller number of models used in this study could lead to a narrower confidence interval than the one reported in D18.

The confidence interval for recovery dates is formed from the predictive uncertainty in the WM from the TSAM (for which the 95 % confidence interval is 2054–2059) and the uncertainty associated with the weighting process. Choices made about which models and metrics to include influence the return dates and therefore introduce uncertainty. This is similar to the concept of an "ensemble of opportunity", which is that only modelling centres with the time, resources or interest take part in certain model ensembles. To quantify this uncertainty, we performed a dropout test where a model and a metric were systematically left out of the recovery date calculation. This was done for all combinations of models ($N = 11$) and metrics ($M = 7$), providing a range of 77 different recovery dates between 2052 and 2058. Combining the TSAM and dropout uncertainties produces a 95 % confidence interval of 2052–2060. We additionally tested dropping out up to three metrics at a time and observed that the confidence interval did not notably increase in size.

Figure ~~2~~ 3 shows the model weights for individual metrics and in total as found using Eqs. (1) and (2). Good agreement is shown between the models for the metrics of lower stratospheric temperature, the temperature gradient, and the TCO gradient. There is one exception of UMUKCA-UCAM which exhibits a colder pole compared to the ensemble and observations. Resultantly, UMUKCA-UCAM is down-weighted for its lower performance at replicating the historic lower stratospheric temperature. Dissimilarity to the rest of the ensemble will contrastingly increase the weighting but to a lesser effect than the down weighting for performance, due in part to the values of the sigma parameters. In spite of a bias in absolute lower stratospheric temperature, UMUKCA-UCAM does reproduce the trend in the lower stratospheric temperature with similar skill to the other models.

Due to the nudging of temperature that takes place in most of the specified dynamics simulations, we would expect stratospheric temperatures to be reasonably well simulated. However, variation exists in nudging methods in addition to inter model differences and this leads to part of the variability in weights (Orbe et al., 2018; Chrysanthou et al., 2019). For the ozone-temperature metrics, which although formed from variables linked to nudged fields are more complex in their construction, we see a much less uniform spread of weights. Furthermore, for processes not directly linked to nudged variables (hydrogen chloride, ozone, and the polar vortex breakdown trend) there is much less agreement between models. This is captured in the weights of these metrics which show just a few models possessing large weights.

The total weighting, formed from the ~~summation~~ mean of individual metric weights per model, is largely influenced by CNRM-CM5-3, which has a weight of 0.27 (297 % the value of a uniform weighting). The CNRM-CM5-3 simulations are more successful at simulating metrics whilst being reasonably independent from other models, leading to a weight with greater prominence than the other models. This does not mean that CNRM-CM5-3 is the most skilful model. For example, if two nearly identical models had the highest performance, their final weights would be much lower as they would be down-weighted for their similarity. All models are contributing towards the weighted ensemble mean providing confidence that our weighting methodology is not over-tuned and returning model weights of zero. The lowest total model weight is ~~55~~ 45 % the value of a uniform weighting.

4.2 Testing the methodology

We performed a perfect model test (Sect. 3.3) to assess the skill of the weighted mean projection, the results of which are shown in Fig. 4. The perfect model test shows that, on average, using this weighting methodology produces a WM which is closer to the 'truth' than the MMM by 1 DU. In addition to improvements in projections, the pseudo recovery dates are better predicted on average, with a maximal improvement of 6 yr.

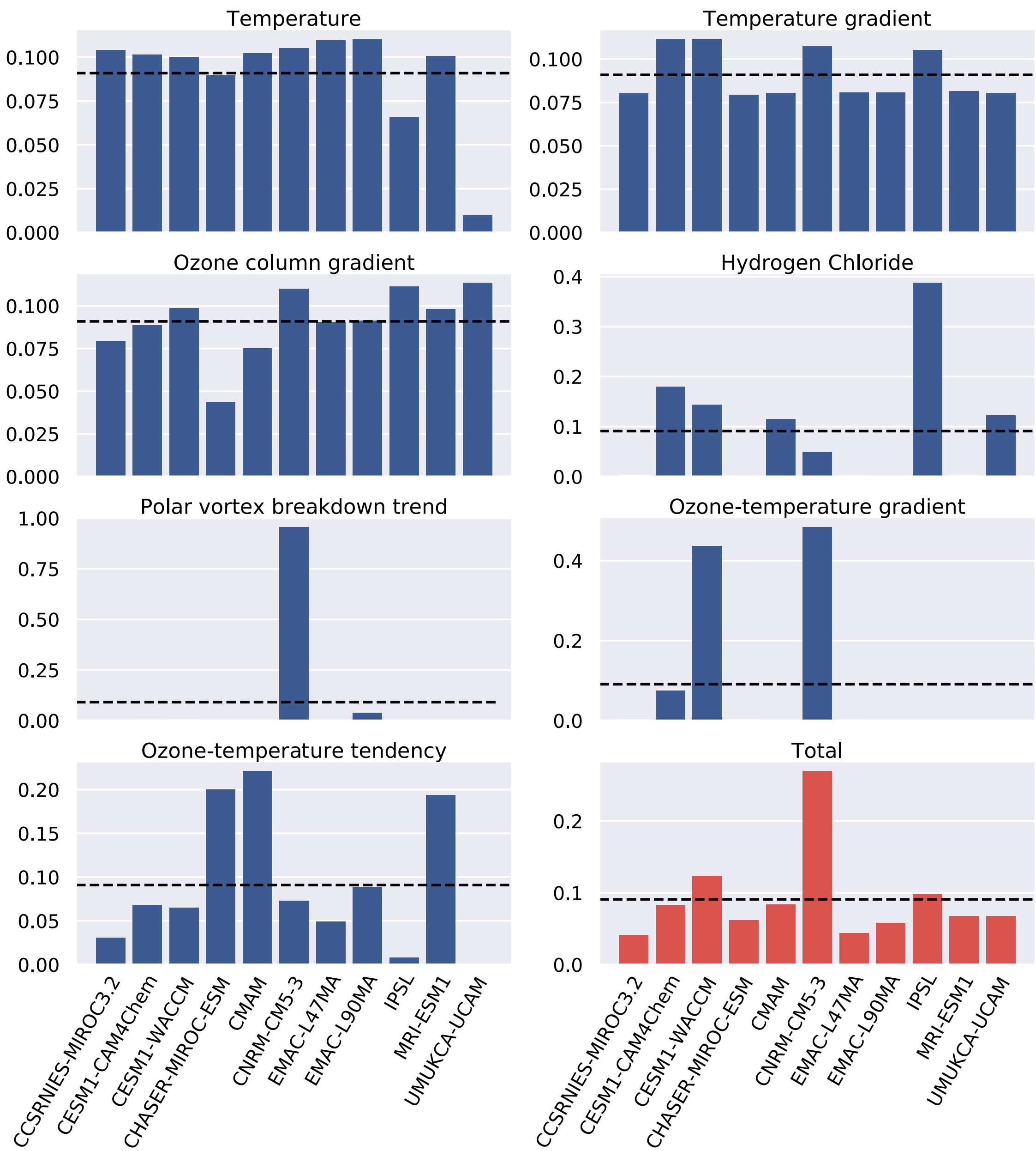


Figure 3. Model weights for each of the seven metrics are all shown in blue. The weights account for both performance and independence and are found using Eq. (1). The total weights, as found from Eq. (2), are shown in red and were the weights used to construct the weighted mean shown in Fig. 2. The black dashed line indicates a uniform weighting as prescribed by a multi model mean.

Three models, when treated as the pseudo truth, do not show an improvement of the WM with respect to the MMM. Note that this is not poor performance of the model in question, rather that the weighting methodology does not do an adequate job of creating a weighted projection for that model as the pseudo truth. Using CHASER-MIROC-ESM as the pseudo truth gives

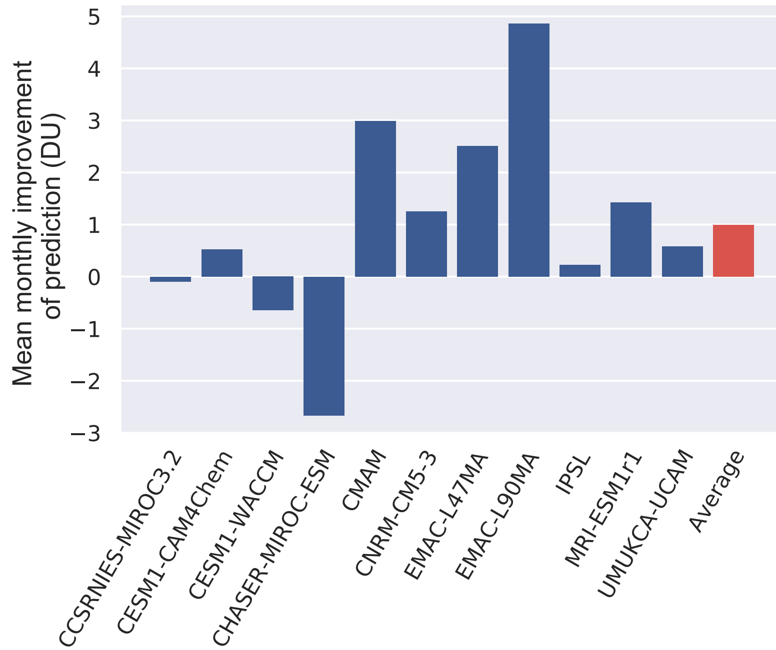


Figure 4. Results of the perfect model test. The ~~average~~ mean monthly improvement in the Antarctic October TCO projection (1960–2095) of the WM compared to the MMM for each model taken as the pseudo truth. The average shown in red is the improvement across all the perfect model tests. No conclusions about overall model skill should be drawn from this plot.

a worse WM projection than if we used the MMM. However, the average correlation between the CHASER-MIROC-ESM-simulated TCO and other models in the ensemble is the lowest at 0.65, compared to the average ensemble cross correlation score of 0.81. Since a weighted mean is a linear combination of models in the ensemble, it is understandable that models with low correlation to CHASER-MIROC-ESM will be less skilful at replicating its TCO time series. This is why an improvement is not seen for CHASER-MIROC-ESM as the pseudo truth in the perfect model testing.

We also performed out-of-sample testing on the WM projection for the years 2010–2016 by comparing it to the TCO observational time series which was smoothed as described in Sect. 3.1 to remove inter annual variability. The root mean squared error (RMSE) was used as the metric for goodness of fit. This range of years is chosen as it is the overlap between the TCO observations, and the years not used in the creation of the weighting. The RMSE of the WM is on average 202 DU^2 less than the MMM per year and the RMSE values were 1510 DU^2 and 2720 DU^2 for the WM and MMM respectively for the out-of-sample period.

4.3 Model independence

The current design of model inter-comparison projects does not account for structural similarities in models, ranging from sharing transport schemes to entire model components. Therefore, a key part of generating an informed weighting is con-

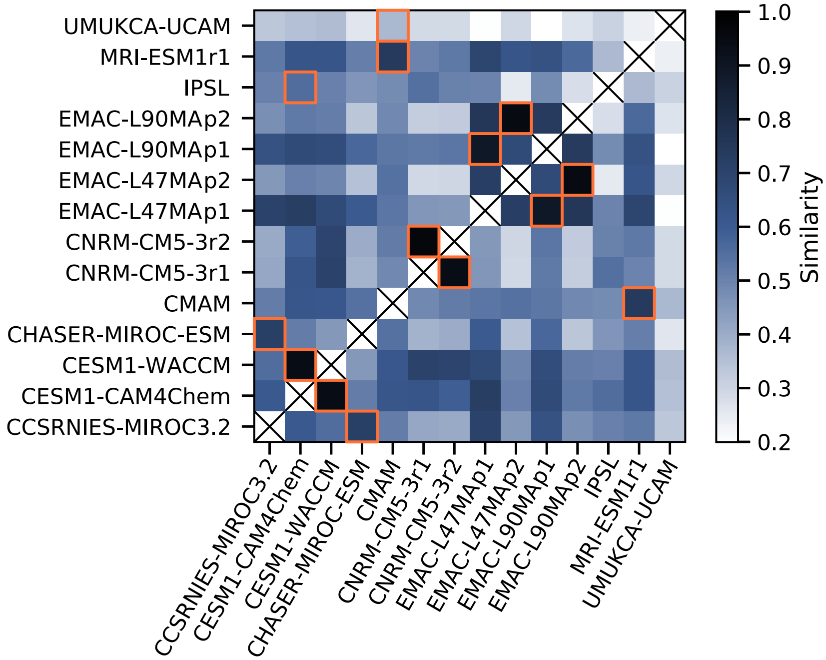


Figure 5. Inter-model similarity across all refC1SD models as calculated by Eq. (3). A similarity of 1 denotes models which are identical for all the metrics, whereas a lower similarity shows a greater independence. The orange boxes highlight the model most similar to the model on the y-axis.

sidering how alike any two models are. The weighting scheme presented here accounts for model independence through the denominator in Eq. (1).

The refC1SD scenario from CCMI consists of 14 different simulations, some of which are with different models, whereas others are just different realisations of the same models. Note that there are more models used here than in the creation of the Antarctic ozone projection. This is because for the weighted projection we require both a refC1SD and a refC2 simulation for each model, but for similarity analysis we can use all the refC1SD simulations. For these model runs we calculated a similarity index s_{ij} (shown in Eq. (3)) which is the similarity between models i and j averaged across all the performance metrics, where n_k is the size of the data for metric k.

$$s_{ij} = \frac{1}{M}\sum_{k=1}^{M} \exp\left(\frac{-S_{ijk}^2}{\sigma_S n_k}\right) \tag{3}$$

Similarities between all refC1SD models are shown in Fig. 5. We also found the maximum value of s_{ij} for each model, indicating the model which model i is most similar to. The most alike models are the two realisations of CNRM-CM5-3, which are the same models running with slightly different initial conditions. We also see high similarity between the two variations of the CESM model, CESM-WACCM and CESM-CAM4Chem. CESM1-CAM4Chem is the low-top version of

CESM1-WACCM, meaning that up to the stratosphere the two models should be much alike (Morgenstern et al., 2017). Analysing the EMAC models this way presents an interesting observation: changing the nudging method, has a greater impact on model similarity than changing the number of vertical levels (the difference between EMAC-L47MAr1i1p1 and EMAC-L47MAr1i1p2, and likewise the 90 level model variant, is that the p1 variant additionally nudges to the global mean temperature (Jöckel et al., 2016)). CHASER-MIROC-ESM and CCSRNIES-MIROC3.2 are two other models which are identified as similar albeit at a lower value. Considering that these two models are built upon the same MIROC general circulation model it is not a surprise that we see a similarity. That the weighting framework can identify all of the models with known similarities (same institution, or realisations) confirms confidence in the methodology and means that we are down-weighting similar models.

5 Discussion

The projection of the ozone hole recovery date presented here makes use of an ensemble of the latest generation of CCMs and a weighting methodology that accounts for complexities within model ensembles. While the ozone recovery date found in this work (2056) is different to that found by Dhomse et al. (2018) (2062), these two dates are not easily comparable as they are created from different subsets of the same ensemble. For our subset of models, the MMM recovery date was 3 years earlier (2053) than the WM. Although the return dates are not significantly different, for the period of peak ozone depletion (especially between 1990 and 2030) the MMM projection is significantly different to the WM. As the model subsets in this work, for the WM and MMM remain the same, the variation in the projections is entirely due to the construction of the WM.

The CNRM-CM5-3 model received the largest weight of 0.27, giving it three times the influence in the WM than in a MMM. Initially this may seem as if we are placing too much importance on one model, but consider that in a standard MMM, a model which runs three simulations with different combinations of components will have three times the influence of a model with a single simulation. Furthermore, CNRM-CM5-3 is not weighted higher because it ran more simulations, it is weighted higher because it is skilful at simulating hindcasts whilst maintaining a level of independence.

Central to the weighting methodology is the selection of metrics requiring expert knowledge. The set of metrics we chose, were grounded in scientific understanding and produce a good improvement of the weighted projection compared to the MMM. There are numerous other metrics of varying complexity which could be considered, such as the size of the ozone hole or the abundance of polar stratospheric clouds. These extra metrics could improve the model weighting and give a more accurate projection, but testing an exhaustive collection of metrics was not our aim, and there are not always appropriate measurements to validate the metrics with. We have shown a weighting framework which improves upon the current methodology for combining model ensembles, and is also flexible and adaptable to which ever metric choices the user deems reasonable. Furthermore, the low range in return dates produced from the dropout testing shows that the results produced in this weighting framework are robust to metric and model choices. This is a desirable effect of a methodology to provide stable results irrespective of fluctuations in the input.

It is reassuring to know that the methodology is robust to metric choices as we are often constrained by the availability of observational data. In this work we benefit from the decades of interest in polar ozone which have led to datasets of a length suitable for constructing model weights. This highlights the importance of continued production of good observational datasets because, although perfect model testing allows us a form of testing which forgoes the need for observations, weighting methodologies must be grounded in some estimate of the truth.

Abramowitz et al. (2019) discuss approaches for assessing model dependence and performance, and mention caveats around the notion of temporal transitivity: is model behaviour comparable between two distinct temporal regions? Here, we rephrase the question to be: are the weights generated from the hindcast scenario relevant and applicable to the forecast scenario? This not only questions temporal transitivity, but also that models may have codified differences between scenarios in addition to differences in physical and chemical regimes. In this study, scenario transitivity (as we call it) is demonstrated through perfect model testing. On average the WM produced a better (closer to the pseudo truth) projection than if we had considered the MMM. This shows that weights calculated from the refC1SD hindcasts produce better projections from the refC2 forecasts and are therefore transitive between the two scenarios.

We generated weights from the refC1SD simulations which means that some metrics we chose are based on nudged variables, such as the lower stratospheric temperature gradient. As a result, one might expect that the model skill for these metrics should be equal, though given Fig. 3 this is not true. One may then expect that the weighting is not capturing model skill, but instead the skill of the models' nudging mechanisms~~. This is harder to test,~~ ; the models are nudged on different timescales ranging from 0.5 hr to 50 hr and from varying reanalysis products (Orbe et al., 2020). We use the perfect model test to show that the utility of the weighting methodology is not compromised by using models with such a variety in nudging time-scales and methods.

~~We justified the use of~~

As the perfect model test produces better projections, for models which are nudged in a variety of ways, we can conclude that the weighting is not dominated by nudging. Take for example UMUKCA-UCAM which is nudged quite differently compared to the ensemble, as evidenced by a southern pole significantly colder than the ensemble. When we take UMUKCA-UCAM as the pseudo truth (temporarily assuming the UMUKCA-UCAM output is the observational truth) we generate weights based upon the refC1SD simulations and test them on the refC2 simulations. The weights generated are based on the dynamical system simulated in refC1SD which includes any model nudging. We can test how well these weights apply to a different dynamical system without nudging (refC2). As we see an improvement in the WM compared to the MMM we can conclude that the weights generated from the refC1SD dynamical system can be applied to the refC2 dynamical system. If there hadn't been an improvement, then the dynamical systems described by refC1SD and refC2 may be too dissimilar for this weighting methodology and the weights may instead have been dominated by how well models are nudged. Nudging may be influencing the weights, but not to a degree that the accuracy of the projection suffers. Orbe et al. (2020) highlight the need for care when using the nudged simulations and we would like any future work on model weighting to quantify the impact of nudging upon model weights to reflect this.

We justified using the nudged refC1SD simulations, despite these considerations, for two reasons. Firstly, ~~that~~ these nudged simulations give the models the best chance at matching the observational record, by providing relatively consistent meteorology across the models. The free running CCMI hindcast simulations (refC1) have a large ensemble variance and, despite producing potentially realistic atmospheric states, are not directly comparable to observational records. Secondly, the perfect model testing discussed above, demonstrates that the nudging doesn't have a ~~negative~~ detrimental effect on the ~~weighting. As the perfect model test produces better projections, for models which are nudged in a variety of ways, we can conclude that the weightingis not measuring nudging. Take for example UMUKCA-UCAM which is nudged quite differently, as evidenced by a colder pole than the ensemble. If the methodology was testing nudging, we would expect the perfect model test, when using UMUKCA-UCAM as the pseudo truth, to not produce a WM projection which was better than the MMM, because the nudging in UMUKCA-UCAM is not like any other model in the ensemble. However, this is not the case, and the WM projection is better than the MMM, confirming that the weighting is not largely dependent on the nudging process.~~model weighting.

Although we were not seeking to grade the CCMs as per Waugh and Eyring (2008), the construction of a weighted mean provides insight into model performance which would not be considered in a MMM. This is of some relevance as the CCMI ensemble has not undergone the same validation as its predecessors, such as CCMVal (Eyring et al., 2008). Additionally, we gain insight into model dependence shown in Sect. 4.2. Whilst this approach may not be as illuminating as Knutti et al. (2013), where they explored the genealogy of CMIP5 models through statistical methods, or Boé (2018), who analysed similarity through model components and version numbers, it successfully identified the known inter-model similarities. More complex methods are desirable, especially those that consider the history of the models' developments. Nevertheless, the simplicity of quantifying inter-model distances as a measure of dependence lends itself well to model weighting.

6 Conclusions

We have presented a model weighting methodology, which considers model dependence and model skill. We applied this over a suite of metrics grounded in scientific understanding to Antarctic ozone depletion and subsequent recovery. In particular we have shown that the weighted projection of the total ozone column trend, with inter-annual variability removed, predicts recovery by 2056 with a 95 % confidence interval of 2052–2060. Through perfect model testing we demonstrated that on average a weighted mean performs better than the current community standard of calculating a multi model mean. Additionally, the perfect model test, a necessary step in validating the methodology, showed a level of transitivity between the free running and the specified dynamics simulations.

This methodology addresses the known shortcomings of an ensemble multi model mean which include, the problem of ensembles including many similar models, and the inability to factor in model performance. It does this by quantifying skill and independence for all models in the ensemble over a selection of metrics which are chosen for their physical relevance to the phenomena of interest. This weighting methodology is still subject to some of the same limitations of taking an ensemble mean: i.e., we are still limited by what the models simulate. For example, in the case of ozone depletion, a weighted mean is no more likely to capture the ozone changes due to the recent fugitive CFC-11 release (Rigby et al., 2019). Instead it allows us to maximise the utility of the ensemble and, provided we are cautious of over-fitting, it allows us to make better projections.

Addressing the shortcomings and presenting possible improvements of methods for averaging model ensembles is timely given the current running of CMIP6 simulations (Eyring et al., 2016). That ensemble could arguably be the largest climate model ensemble created to date, in terms of the breadth of models considered. Therefore, the need for tools to analyse vast swathes of data efficiently for multiple interests is still growing. The models within CMIP6 are likely not all independent, which could affect the robustness of results from the ensemble by biasing the output towards groups of similar models. The similarity analysis within this work would allow users of the ensemble data to understand if ensemble biases are emerging from similar models and acknowledge how this may impact their results.

In summary, we have presented a flexible and useful methodology, which has applications throughout the environmental sciences. It is not a silver bullet for creating the perfect projection for all circumstances; however, it can be used to construct a phenomenon-specific analysis process that can account for model skill and model independence, both of which ~~will~~ can improve ensemble projections compared to a multi-model mean.

Code and data availability. The jupyter notebook used to run the analysis, along with a collection of functions to produce weightings from ensembles, can be found at https://doi.org/10.5281/zenodo.3624522. The CCMI model output was retrieved from the Centre for Environmental Data Analysis (CEDA), the Natural Environment Research Council's Data Repository for Atmospheric Science and Earth Observation (http://data.ceda.ac.uk/badc/wcrp-ccmi/data/CCMI-1/output), and from NCAR's Climate Data Gateway (http://www.earthsystemgrid.org).

Author contributions. MA developed the methods and led the analysis, and conceived the study alongside JSH and PJY, who made major contributions as the work progressed. MA drafted the manuscript, with the guidance of PJY and JSH and input from JFL. JFL and all the other co-authors provided model simulation data. All the co-authors provided model output and helped with finalising the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was supported by the Natural Environment Research Council [NERC grant reference number NE/L002604/1], with Matt Amos's studentship through the ENVISION Doctoral Training Partnership. Paul J Young is partially supported by the Data Science of the Natural Environment (DSNE) project, funded by the UK Engineering and Physical Sciences Research Council (EPSRC; grant number EP/R01860X/1) as well as the Impact of Short-Lived Halocarbons on Ozone and Climate (ISHOC) project, funded by the UK Natural Environment Research Council (NERC; grant number: NE/R004927/1). We would like to thank Bodeker Scientific, funded by the New Zealand Deep South National Science Challenge, for providing the combined NIWA-BS total column ozone database. We also acknowledge the GOZCARDs team for the production of the HCl record and Remote Sensing Systems for the MSU TLS record. We acknowledge the modelling groups for making their simulations available for this analysis, the joint WCRP SPARC/IGAC Chemistry-Climate Model Initiative (CCMI) for organising and coordinating the model data analysis activity, and the British Atmospheric Data Centre (BADC) for collecting and archiving the CCMI model output. The EMAC simulations have been performed at the German Climate Computing Centre (DKRZ) through support from the Bundesministerium für Bildung und Forschung (BMBF). DKRZ and its scientific steering committee are gratefully acknowledged for providing the HPC and data archiving resources for this consortial project ESCiMo (Earth System Chemistry integrated Modelling). The CCSRNIES-MIROC3.2 model computations were performed on NEC-SX9/A(ECO) and NEC SX-ACE computers at the CGER, NIES, and supported by the Environment Research and Technology Development Funds of the Ministry of the Environment, Japan (2-1303) and Environment Restoration and Conservation Agency, Japan (2-1709).

References

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A.: ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, Earth System Dynamics, 10, 91–105, https://doi.org/10.5194/esd-10-91-2019, 2019.

Akiyoshi, H., Nakamura, T., Miyasaka, T., Shiotani, M., and Suzuki, M.: A nudged chemistry-climate model simulation of chemical constituent distribution at northern high-latitude stratosphere observed by SMILES and MLS during the 2009/2010 stratospheric sudden warming, Journal of Geophysical Research: Atmospheres, 121, 1361–1380, https://doi.org/10.1002/2015JD023334, 2016.

Ball, W. T., Alsing, J., Mortlock, D. J., Staehelin, J., Haigh, J. D., Peter, T., Tummon, F., Stübi, R., Stenke, A., Anderson, J., Bourassa, A., Davis, S. M., Degenstein, D., Frith, S., Froidevaux, L., Roth, C., Sofieva, V., Wang, R., Wild, J., Yu, P., Ziemke, J. R., and Rozanov, E. V.: Evidence for a continuous decline in lower stratospheric ozone offsetting ozone layer recovery, Atmospheric Chemistry and Physics, 18, 1379–1394, https://doi.org/10.5194/acp-18-1379-2018, 2018.

Bednarz, E. M., Maycock, A. C., Abraham, N. L., Braesicke, P., Dessens, O., and Pyle, J. A.: Future Arctic ozone recovery: the importance of chemistry and dynamics, Atmospheric Chemistry and Physics, 16, 12 159–12 176, https://doi.org/10.5194/acp-16-12159-2016, 2016.

Bellenger, H., Guilyardi, E., Leloup, J., Lengaigne, M., and Vialard, J.: ENSO representation in climate models: from CMIP3 to CMIP5, Climate Dynamics, 42, 1999–2018, https://doi.org/10.1007/s00382-013-1783-z, 2014.

Berrisford, P., Dee, D., Poli, P., Brugge, R., Fielding, M., Fuentes, M., Kållberg, P., Kobayashi, S., Uppala, S., and Simmons, A.: The ERA-Interim archive Version 2.0, p. 23, 2011.

Bodeker, G. E., Shiona, H., and Eskes, H.: Indicators of Antarctic ozone depletion, Atmospheric Chemistry and Physics, 5, 2603–2615, https://doi.org/10.5194/acp-5-2603-2005, 2005.

Bodeker, G. E., Nitzbon, J., Lewis, J., Schwertheim, A., and Tradowsky, J. S.: NIWA-BS Total Column Ozone Database, https://doi.org/10.5281/zenodo.1346424, https://doi.org/10.5281/zenodo.1346424, 2018.

Boé, J.: Interdependency in Multimodel Climate Projections: Component Replication and Result Similarity, Geophysical Research Letters, 45, 2771–2779, https://doi.org/10.1002/2017GL076829, 2018.

Brunner, L., Lorenz, R., Zumwald, M., and Knutti, R.: Quantifying uncertainty in European climate projections using combined performance-independence weighting, Environmental Research Letters, 14, 124 010, https://doi.org/10.1088/1748-9326/ab492f, 2019.

Butchart, N., Cionni, I., Eyring, V., Shepherd, T. G., Waugh, D. W., Akiyoshi, H., Austin, J., Brühl, C., Chipperfield, M. P., Cordero, E., Dameris, M., Deckert, R., Dhomse, S., Frith, S. M., Garcia, R. R., Gettelman, A., Giorgetta, M. A., Kinnison, D. E., Li, F., Mancini, E., McLandress, C., Pawson, S., Pitari, G., Plummer, D. A., Rozanov, E., Sassi, F., Scinocca, J. F., Shibata, K., Steil, B., and Tian, W.: Chemistry–Climate Model Simulations of Twenty-First Century Stratospheric Climate and Circulation Changes, Journal of Climate, 23, 5349–5374, https://doi.org/10.1175/2010JCLI3404.1, 2010.

Butler, A., Daniel, J. S., Portmann, R. W., Ravishankara, A., Young, P. J., Fahey, D. W., and Rosenlof, K. H.: Diverse policy implications for future ozone and surface UV in a changing climate, Environmental Research Letters, 11, 064 017, https://doi.org/10.1088/1748-9326/11/6/064017, 2016.

Chipperfield, M. P., Bekki, S., Dhomse, S., Harris, N. R., Hassler, B., Hossaini, R., Steinbrecht, W., Thiéblemont, R., and Weber, M.: Detecting recovery of the stratospheric ozone layer, Nature, 549, 211, https://doi.org/https://doi.org/10.1038/nature23681, 2017.

Christensen, J. H., Kjellström, E., Giorgi, F., Lenderink, G., and Rummukainen, M.: Weight assignment in regional climate models, Climate Research, 44, 179–194, https://doi.org/10.3354/cr00916, 2010.

Chrysanthou, A., Maycock, A. C., Chipperfield, M. P., Dhomse, S., Garny, H., Kinnison, D., Akiyoshi, H., Deushi, M., Garcia, R. R., Jöckel, P., Kirner, O., Pitari, G., Plummer, D. A., Revell, L., Rozanov, E., Stenke, A., Tanaka, T. Y., Visioni, D., and Yamashita, Y.: The effect of atmospheric nudging on the stratospheric residual circulation in chemistry–climate models, Atmospheric Chemistry and Physics, 19, 11 559–11 586, https://doi.org/10.5194/acp-19-11559-2019, 2019.

Claxton, T., Hossaini, R., Wild, O., Chipperfield, M. P., and Wilson, C.: On the Regional and Seasonal Ozone Depletion Potential of Chlorinated Very Short-Lived Substances, Geophysical Research Letters, 46, 5489–5498, https://doi.org/https://doi.org/10.1029/2018GL081455, 2019.

Deushi, M. and Shibata, K.: Development of a Meteorological Research Institute chemistry-climate model version 2 for the study of tropospheric and stratospheric chemistry, Papers in Meteorology and Geophysics, 62, 1–46, https://doi.org/10.2467/mripapers.62.1, 2011.

Dhomse, S. S., Kinnison, D., Chipperfield, M. P., Salawitch, R. J., Cionni, I., Hegglin, M. I., Abraham, N. L., Akiyoshi, H., Archibald, A. T., Bednarz, E. M., et al.: Estimates of ozone return dates from Chemistry-Climate Model Initiative simulations, Atmospheric Chemistry and Physics, 18, 8409–8438, https://doi.org/10.5194/acp-18-8409-2018, 2018.

Dufresne, J.-L., Foujols, M.-A., Denvil, S., Caubel, A., Marti, O., Aumont, O., Balkanski, Y., Bekki, S., Bellenger, H., Benshila, R., et al.: Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5, Climate Dynamics, 40, 2123–2165, https://doi.org/10.1007/s00382-012-1636-1, 2013.

Eyring, V., Butchart, N., Waugh, D. W., Akiyoshi, H., Austin, J., Bekki, S., Bodeker, G. E., Boville, B. A., Brühl, C., Chipperfield, M. P., Cordero, E., Dameris, M., Deushi, M., Fioletov, V. E., Frith, S. M., Garcia, R. R., Gettelman, A., Giorgetta, M. A., Grewe, V., Jourdain, L., Kinnison, D. E., Mancini, E., Manzini, E., Marchand, M., Marsh, D. R., Nagashima, T., Newman, P. A., Nielsen, J. E., Pawson, S., Pitari, G., Plummer, D. A., Rozanov, E., Schraner, M., Shepherd, T. G., Shibata, K., Stolarski, R. S., Struthers, H., Tian, W., and Yoshiki, M.: Assessment of temperature, trace species, and ozone in chemistry-climate model simulations of the recent past, Journal of Geophysical Research: Atmospheres, 111, https://doi.org/10.1029/2006JD007327, 2006.

Eyring, V., Chipperfield, M. P., Giorgetta, M. A., Kinnison, D. E., Manzini, E., Matthes, K., Newman, P. A., Pawson, S., Shepherd, T. G., and Waugh, D. W.: Overview of the new CCMVal reference and sensitivity simulations in support of upcoming ozone and climate assessments and the planned SPARC CCMVal report, SPARC Newsletter, 30, 20–26, http://oceanrep.geomar.de/15163/, 2008.

Eyring, V., Cionni, I., Bodeker, G. E., Charlton-Perez, A., Kinnison, D. E., Scinocca, J. F., Waugh, D. W., Akiyoshi, H., Bekki, S., Chipperfield, M. P., et al.: Multi-model assessment of stratospheric ozone return dates and ozone recovery in CCMVal-2 models, Atmospheric Chemistry and Physics, 10, 9451–9472, https://doi.org/10.5194/acp-10-9451-2010, 2010.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., et al.: Evaluation of Climate Models, in: Climate Change 2013 – The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, p. 741–866, Cambridge University Press, https://doi.org/10.1017/CBO9781107415324.020, 2014.

Froidevaux, L., Anderson, J., Wang, H.-J., Fuller, R. A., Schwartz, M. J., Santee, M. L., Livesey, N. J., Pumphrey, H. C., Bernath, P. F., Russell III, J. M., and McCormick, M. P.: Global OZone Chemistry And Related trace gas Data records for the Stratosphere (GOZCARDS): methodology and sample results with a focus on HCl, H_2O, and O_3, Atmospheric Chemistry and Physics, 15, 10 471–10 507, https://doi.org/10.5194/acp-15-10471-2015, 2015.

Garcia, R. R., Smith, A. K., Kinnison, D. E., Cámara, Á. d. l., and Murphy, D. J.: Modification of the gravity wave parameterization in the Whole Atmosphere Community Climate Model: Motivation and results, Journal of the Atmospheric Sciences, 74, 275–291, https://doi.org/10.1175/JAS-D-16-0104.1, 2017.

Gillett, N. P.: Weighting climate model projections using observational constraints, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 373, 20140 425, https://doi.org/https://doi.org/10.1098/rsta.2014.0425, 2015.

Giorgi, F. and Mearns, L. O.: Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging"(REA) method, Journal of Climate, 15, 1141–1158, https://doi.org/10.1175/1520-0442(2002)015<1141:COAURA>2.0.CO;2, 2002.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, Journal of Geophysical Research: Atmospheres, 113, https://doi.org/10.1029/2007JD008972, 2008.

Harrison, S. P., Bartlein, P., Izumi, K., Li, G., Annan, J., Hargreaves, J., Braconnot, P., and Kageyama, M.: Evaluation of CMIP5 palaeo-simulations to improve climate projections, Nature Climate Change, 5, 735, https://doi.org/10.1038/nclimate2649, 2015.

Haughton, N., Abramowitz, G., Pitman, A., and Phipps, S. J.: Weighting climate model ensembles for mean and variance estimates, Climate dynamics, 45, 3169–3181, https://doi.org/doi.org/10.1007/s00382-015-2531-3, 2015.

Herger, N., Abramowitz, G., Sherwood, S., Knutti, R., Angélil, O., and Sisson, S. A.: Ensemble optimisation, multiple constraints and over-confidence: a case study with future Australian precipitation change, Climate Dynamics, 53, 1581–1596, https://doi.org/10.1007/s00382-019-04690-8, 2019.

Hossaini, R., Atlas, E., Dhomse, S. S., Chipperfield, M. P., Bernath, P. F., Fernando, A. M., Mühle, J., Leeson, A. A., Montzka, S. A., Feng, W., et al.: Recent trends in stratospheric chlorine from very short-lived substances, Journal of Geophysical Research: Atmospheres, 124, 2318–2335, https://doi.org/10.1029/2018JD029400, 2019.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J. C., Balaji, V., Duan, Q., Folini, D., Ji, D., Klocke, D., Qian, Y., Rauser, F., Rio, C., Tomassini, L., Watanabe, M., and Williamson, D.: The art and science of climate model tuning, Bulletin of the American Meteorological Society, 98, 589–602, https://doi.org/10.1175/BAMS-D-15-00135.1, 2017.

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

Iglesias-Suarez, F., Young, P. J., and Wild, O.: Stratospheric ozone change and related climate impacts over 1850–2100 as modelled by the ACCMIP ensemble, Atmospheric Chemistry and Physics, 16, 343–363, https://doi.org/10.5194/acp-16-343-2016, 2016.

Imai, K., Manago, N., Mitsuda, C., Naito, Y., Nishimoto, E., Sakazaki, T., Fujiwara, M., Froidevaux, L., von Clarmann, T., Stiller, G. P., et al.: Validation of ozone data from the Superconducting Submillimeter-Wave Limb-Emission Sounder (SMILES), Journal of Geophysical Research: Atmospheres, 118, 5750–5769, https://doi.org/10.1002/jgrd.50434, 2013.

Jöckel, P., Kerkweg, A., Pozzer, A., Sander, R., Tost, H., Riede, H., Baumgaertner, A., Gromov, S., and Kern, B.: Development cycle 2 of the modular earth submodel system (MESSy2), Geoscientific Model Development, 3, 717–752, https://doi.org/10.5194/gmd-3-717-2010, 2010.

Jöckel, P., Tost, H., Pozzer, A., Kunze, M., Kirner, O., Brenninkmeijer, C. A. M., Brinkop, S., Cai, D. S., Dyroff, C., Eckstein, J., Frank, F., Garny, H., Gottschaldt, K.-D., Graf, P., Grewe, V., Kerkweg, A., Kern, B., Matthes, S., Mertens, M., Meul, S., Neumaier, M., Nützel, M., Oberländer-Hayn, S., Ruhnke, R., Runde, T., Sander, R., Scharffe, D., and Zahn, A.: Earth System Chemistry integrated Modelling (ESCiMo) with the Modular Earth Submodel System (MESSy) version 2.51, Geoscientific Model Development, 9, 1153–1200, https://doi.org/10.5194/gmd-9-1153-2016, 2016.

Jonsson, A., De Grandpre, J., Fomichev, V., McConnell, J., and Beagley, S.: Doubled CO_2-induced cooling in the middle atmosphere: Photochemical analysis of the ozone radiative feedback, Journal of Geophysical Research: Atmospheres, 109, https://doi.org/10.1029/2004JD005093, 2004.

Knutti, R.: The end of model democracy?, Climatic Change, 102, 395–404, https://doi.org/10.1007/s10584-010-9800-2, 2010.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A.: Challenges in Combining Projections from Multiple Climate Models, Journal of Climate, 23, 2739–2758, https://doi.org/10.1175/2009JCLI3361.1, 2010.

Knutti, R., Masson, D., and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, Geophysical Research Letters, 40, 1194–1199, https://doi.org/10.1002/grl.50256, 2013.

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V.: A climate model projection weighting scheme accounting for performance and interdependence, Geophysical Research Letters, 44, 1909–1918, https://doi.org/10.1002/2016GL072012, 2017.

Lamarque, J.-F., Bond, T. C., Eyring, V., Granier, C., Heil, A., Klimont, Z., Lee, D., Liousse, C., Mieville, A., Owen, B., Schultz, M. G., Shindell, D., Smith, S. J., Stehfest, E., Van Aardenne, J., Cooper, O. R., Kainuma, M., Mahowald, N., McConnell, J. R., Naik, V., Riahi, K., and van Vuuren, D. P.: Historical (1850-2000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: methodology and application, Atmospheric Chemistry and Physics, 10, 7017–7039, https://doi.org/10.5194/acp-10-7017-2010, 2010.

Lamarque, J.-F., Shindell, D. T., Josse, B., Young, P., Cionni, I., Eyring, V., Bergmann, D., Cameron-Smith, P., Collins, W. J., Doherty, R., et al.: The Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP): overview and description of models, simulations and climate diagnostics, Geoscientific Model Development, 6, 179–206, https://doi.org/doi.org/10.5194/gmd-6-179-2013, 2013.

Langematz, U., Tully, M., Calvo, N., Dameris, M., de Laat A.T.J, Klekociuk, A., Muller, R., Young, P., et al.: Polar stratospheric ozone: past, present, and future, in: Scientific Assessment of Ozone Depletion: 2018, Global Ozone Research and Monitoring Project-Report No. 58, WMO, 2018.

Lee, M., Jun, M., and Genton, M. G.: Validation of CMIP5 multimodel ensembles through the smoothness of climate variables, Tellus, Series A: Dynamic Meteorology and Oceanography, 67, https://doi.org/10.3402/tellusa.v67.23880, 2015.

Lorenz, R., Herger, N., Sedláček, J., Eyring, V., Fischer, E. M., and Knutti, R.: Prospects and Caveats of Weighting Climate Models for Summer Maximum Temperature Projections Over North America, Journal of Geophysical Research: Atmospheres, 123, 4509–4526, https://doi.org/10.1029/2017JD027992, 2018.

Marchand, M., Keckhut, P., Lefebvre, S., Claud, C., Cugnet, D., Hauchecorne, A., Lefèvre, F., Lefebvre, M.-P., Jumelet, J., Lott, F., et al.: Dynamical amplification of the stratospheric solar response simulated with the Chemistry-Climate model LMDz-Reprobus, Journal of Atmospheric and Solar-Terrestrial Physics, 75, 147–160, https://doi.org/10.1016/j.jastp.2011.11.008, 2012.

Marsh, D. R., Mills, M. J., Kinnison, D. E., Lamarque, J.-F., Calvo, N., and Polvani, L. M.: Climate change from 1850 to 2005 simulated in CESM1 (WACCM), Journal of climate, 26, 7372–7391, https://doi.org/10.1175/JCLI-D-12-00558.1, 2013.

Mears, C. A. and Wentz, F. J.: Construction of the Remote Sensing Systems V3.2 Atmospheric Temperature Records from the MSU and AMSU Microwave Sounders, Journal of Atmospheric and Oceanic Technology, 26, 1040–1056, https://doi.org/10.1175/2008JTECHA1176.1, 2009.

Michou, M., Saint-Martin, D., Teyssedre, H., Alias, A., Karcher, F., Olivié, D., Voldoire, A., Josse, B., Peuch, V.-H., Clark, H., et al.: A new version of the CNRM Chemistry-Climate Model, CNRM-CCM: description and improvements from the CCMVal-2 simulations, Geoscientific Model Development, 4, 873–900, https://doi.org/10.5194/gmd-4-873-2011, 2011.

Montzka, S. A., Dutton, G. S., Yu, P., Ray, E., Portmann, R. W., Daniel, J. S., Kuijpers, L., Hall, B. D., Mondeel, D., Siso, C., et al.: An unexpected and persistent increase in global emissions of ozone-depleting CFC-11, Nature, 557, 413, https://doi.org/10.1038/s41586-018-0106-2, 2018.

Morgenstern, O., Braesicke, P., O'Connor, F., Bushell, A., Johnson, C., Osprey, S., and Pyle, J.: Evaluation of the new UKCA climate-composition model-Part 1: The stratosphere, Geoscientific Model Development, 2, https://doi.org/10.5194/gmd-2-43-2009, 2009.

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), Geoscientific Model Development, 10, 639–671, https://doi.org/10.5194/gmd-10-639-2017, 2017.

Orbe, C., Yang, H., Waugh, D. W., Zeng, G., Morgenstern, O., Kinnison, D. E., Lamarque, J.-F., Tilmes, S., Plummer, D. A., Scinocca, J. F., et al.: Large-scale tropospheric transport in the Chemistry–Climate Model Initiative (CCMI) simulations, Atmospheric Chemistry and Physics, 18, 7217–7235, https://doi.org/10.5194/acp-18-7217-2018, 2018.

Orbe, C., Plummer, D. A., Waugh, D. W., Yang, H., Jöckel, P., Kinnison, D. E., Josse, B., Marecal, V., Deushi, M., Abraham, N. L., Archibald, A. T., Chipperfield, M. P., Dhomse, S., Feng, W., and Bekki, S.: Description and Evaluation of the specified-dynamics experiment in the Chemistry-Climate Model Initiative, Atmospheric Chemistry and Physics, 20, 3809–3840, https://doi.org/10.5194/acp-20-3809-2020, 2020.

Perlwitz, J., Pawson, S., Fogt, R. L., Nielsen, J. E., and Neff, W. D.: Impact of stratospheric ozone hole recovery on Antarctic climate, Geophysical Research Letters, 35, https://doi.org/10.1029/2008GL033317, 2008.

Pincus, R., Batstone, C. P., Hofmann, R. J. P., Taylor, K. E., and Glecker, P. J.: Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models, Journal of Geophysical Research: Atmospheres, 113, https://doi.org/10.1029/2007JD009334, 2008.

Portmann, R., Daniel, J., and Ravishankara, A.: Stratospheric ozone depletion due to nitrous oxide: influences of other gases, Philosophical Transactions of the Royal Society B: Biological Sciences, 367, 1256–1264, https://doi.org/10.1098/rstb.2011.0377, 2012.

Räisänen, J., Ruokolainen, L., and Ylhäisi, J.: Weighting of model results for improving best estimates of climate change, Climate Dynamics, 35, 407–422, https://doi.org/10.1007/s00382-009-0659-8, 2010.

Reichler, T. and Kim, J.: How Well Do Coupled Models Simulate Today's Climate?, Bulletin of the American Meteorological Society, 89, 303–312, https://doi.org/10.1175/BAMS-89-3-303, 2008.

Rigby, M., Park, S., Saito, T., Western, L., Redington, A., Fang, X., Henne, S., Manning, A., Prinn, R., Dutton, G., et al.: Increase in CFC-11 emissions from eastern China based on atmospheric observations, Nature, 569, 546, https://doi.org/10.1038/s41586-019-1193-4, 2019.

Rybka, H. and Tost, H.: Uncertainties in future climate predictions due to convection parameterisations, Atmospheric Chemistry and Physics, 14, 5561–5576, https://doi.org/10.5194/acp-14-5561-2014, 2014.

Sanderson, B. M., Knutti, R., and Caldwell, P.: Addressing Interdependency in a Multimodel Ensemble by Interpolation of Model Properties, Journal of Climate, 28, 5150–5170, https://doi.org/10.1175/JCLI-D-14-00361.1, 2015a.

Sanderson, B. M., Knutti, R., and Caldwell, P.: A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble, Journal of Climate, 28, 5171–5194, https://doi.org/10.1175/JCLI-D-14-00362.1, 2015b.

Sanderson, B. M., Wehner, M., and Knutti, R.: Skill and independence weighting for multi-model assessments, Geoscientific Model Development, 10, 2379–2395, https://doi.org/10.5194/gmd-10-2379-2017, 2017.

Scinocca, J., McFarlane, N., Lazare, M., Li, J., and Plummer, D.: The CCCma third generation AGCM and its extension into the middle atmosphere, Atmospheric Chemistry and Physics, 8, 7055–7074, https://doi.org/10.5194/acp-8-7055-2008, 2008.

Scinocca, J. F., Stephenson, D. B., Bailey, T. C., and Austin, J.: Estimates of past and future ozone trends from multimodel simulations using a flexible smoothing spline methodology, Journal of Geophysical Research: Atmospheres, 115, https://doi.org/10.1029/2009JD013622, 2010.

Sekiya, T. and Sudo, K.: Role of meteorological variability in global tropospheric ozone during 1970–2008, Journal of Geophysical Research: Atmospheres, 117, https://doi.org/10.1029/2012JD018054, 2012.

Sekiya, T. and Sudo, K.: Roles of transport and chemistry processes in global ozone change on interannual and multidecadal time scales, Journal of Geophysical Research: Atmospheres, 119, 4903–4921, https://doi.org/10.1002/2013JD020838, 2014.

Solomon, S.: Stratospheric ozone depletion: A review of concepts and history, Reviews of Geophysics, 37, 275–316, https://doi.org/10.1029/1999RG900008, 1999.

Solomon, S., Qin, D., Manning, M., Averyt, K., and Marquis, M.: Climate change 2007-the physical science basis: Working group I contribution to the fourth assessment report of the IPCC, vol. 4, Cambridge university press, 2007.

Solomon, S., Kinnison, D., Bandoro, J., and Garcia, R.: Simulation of polar ozone depletion: An update, Journal of Geophysical Research: Atmospheres, 120, 7958–7974, https://doi.org/10.1002/2015JD023365, 2015.

Solomon, S., Ivy, D. J., Kinnison, D., Mills, M. J., Neely, R. R., and Schmidt, A.: Emergence of healing in the Antarctic ozone layer, Science, 353, 269–274, https://doi.org/10.1126/science.aae0061, 2016.

Son, S.-W., Polvani, L. M., Waugh, D. W., Akiyoshi, H., Garcia, R., Kinnison, D., Pawson, S., Rozanov, E., Shepherd, T. G., and Shibata, K.: The Impact of Stratospheric Ozone Recovery on the Southern Hemisphere Westerly Jet, Science, 320, 1486–1489, https://doi.org/10.1126/science.1155939, 2008.

Sudo, K. and Akimoto, H.: Global source attribution of tropospheric ozone: Long-range transport from various source regions, Journal of Geophysical Research: Atmospheres, 112, https://doi.org/10.1029/2006JD007992, 2007.

Sudo, K., Takahashi, M., Kurokawa, J.-i., and Akimoto, H.: CHASER: A global chemical model of the troposphere 1. Model description, Journal of Geophysical Research: Atmospheres, 107, ACH–7, https://doi.org/10.1029/2001JD001113, 2002.

Szopa, S., Balkanski, Y., Schulz, M., Bekki, S., Cugnet, D., Fortems-Cheiney, A., Turquety, S., Cozic, A., Déandreis, C., Hauglustaine, D., et al.: Aerosol and ozone changes as forcing for climate evolution between 1850 and 2100, Climate dynamics, 40, 2223–2250, https://doi.org/10.1007/s00382-012-1408-y, 2013.

Tebaldi, C. and Knutti, R.: The use of the multi-model ensemble in probabilistic climate projections, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 365, 2053–2075, https://doi.org/10.1098/rsta.2007.2076, 2007.

Tebaldi, C., Smith, R. L., Nychka, D., and Mearns, L. O.: Quantifying Uncertainty in Projections of Regional Climate Change: A Bayesian Approach to the Analysis of Multimodel Ensembles, Journal of Climate, 18, 1524–1540, https://doi.org/10.1175/JCLI3363.1, 2005.

Thompson, D. W. and Solomon, S.: Interpretation of recent Southern Hemisphere climate change, Science, 296, 895–899, https://doi.org/10.1126/science.1069270, 2002.

Thompson, D. W. and Solomon, S.: Understanding recent stratospheric climate change, Journal of Climate, 22, 1934–1943, https://doi.org/10.1175/2008JCLI2482.1, 2008.

Thompson, D. W., Solomon, S., Kushner, P. J., England, M. H., Grise, K. M., and Karoly, D. J.: Signatures of the Antarctic ozone hole in Southern Hemisphere surface climate change, Nature geoscience, 4, 741, https://doi.org/10.1038/ngeo1296, 2011.

Tilmes, S., Lamarque, J.-F., Emmons, L., Kinnison, D., Ma, P., Liu, X., Ghan, S., Bardeen, C., Arnold, S., Deeter, M., et al.: Description and evaluation of tropospheric chemistry and aerosols in the Community Earth System Model (CESM1. 2), Geoscientific Model Development, 8, 1395–1426, https://doi.org/10.5194/gmd-8-1395-2015, 2015.

van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G. C., Kram, T., Krey, V., Lamarque, J.-F., Masui, T., Meinshausen, M., Nakicenovic, N., Smith, S. J., and Rose, S. K.: The representative concentration pathways: an overview, Climatic Change, 109, 5, https://doi.org/10.1007/s10584-011-0148-z, 2011.

Voldoire, A., Sanchez-Gomez, E., y Mélia, D. S., Decharme, B., Cassou, C., Sénési, S., Valcke, S., Beau, I., Alias, A., Chevallier, M., et al.: The CNRM-CM5. 1 global climate model: description and basic evaluation, Climate Dynamics, 40, 2091–2121, https://doi.org/10.1007/s00382-011-1259-y, 2013.

Watanabe, S., Hajima, T., Sudo, K., Nagashima, T., Takemura, T., Okajima, H., Nozawa, T., Kawase, H., Abe, M., Yokohata, T., et al.: MIROC-ESM 2010: Model description and basic results of CMIP5-20c3m experiments, Geoscientific Model Development, 4, 845, https://doi.org/10.5194/gmd-4-845-2011, 2011.

Waugh, D. W. and Eyring, V.: Quantitative performance metrics for stratospheric-resolving chemistry-climate models, Atmospheric Chemistry and Physics, 8, 5699–5713, https://doi.org/10.5194/acp-8-5699-2008, 2008.

WMO: Scientific Assessment of Ozone Depletion: 2010, Global Ozone Research and Monitoring Project-Report No. 52, WMO, Geneva, Switzerland, 2011.

WMO: Scientific Assessment of Ozone Depletion: 2018, Global Ozone Research and Monitoring Project-Report No. 58, WMO, Geneva, Switzerland, 2018.

Young, P. J., Rosenlof, K. H., Solomon, S., Sherwood, S. C., Fu, Q., and Lamarque, J.-F.: Changes in stratospheric temperatures and their implications for changes in the Brewer–Dobson circulation, 1979–2005, Journal of Climate, 25, 1759–1772, https://doi.org/10.1175/2011JCLI4048.1, 2012.

Young, P. J., Butler, A. H., Calvo, N., Haimberger, L., Kushner, P. J., Marsh, D. R., Randel, W. J., and Rosenlof, K. H.: Agreement in late twentieth century Southern Hemisphere stratospheric temperature trends in observations and CCMVal-2, CMIP3, and CMIP5 models, Journal of Geophysical Research: Atmospheres, 118, 605–613, https://doi.org/10.1002/jgrd.50126, 2013.

Young, P. J., Naik, V., Fiore, A. M., Gaudel, A., Guo, J., Lin, M., Neu, J., Parrish, D., Reider, H., Schnell, J., et al.: Tropospheric Ozone Assessment Report: Assessment of global-scale model performance for global and regional ozone distributions, variability, and trends, Elementa: Science of the Anthropocene, 6, https://doi.org/10.1525/elementa.265, 2018.

Yuan, X. and Wood, E. F.: On the clustering of climate models in ensemble seasonal forecasting, Geophysical Research Letters, 39, https://doi.org/10.1029/2012GL052735, 2012.

Yukimoto, S.: Meteorological research institute earth system model version 1 (MRI-ESM1): model description, Technical reports of the Meteorological Research Institute, https://doi.org/10.11483/mritechrepo.64, 2011.

Yukimoto, S., Adachi, Y., Hosaka, M., Sakami, T., Yoshimura, H., Hirabara, M., Tanaka, T. Y., Shindo, E., Tsujino, H., Deushi, M., et al.: A new global climate model of the Meteorological Research Institute: MRI-CGCM3—model description and basic performance, Journal of the Meteorological Society of Japan. Ser. II, 90, 23–64, https://doi.org/10.2151/jmsj.2012-A02, 2012.