In our response, reviewer comments are marked in bold, our responses and original text in plain text, and altered text in the paper in bold italic. Additionally we highlight altered text in the tracked changes pdf as requested.

Response to reviewer 1 (J. Mülmenstädt)

We thank the reviewer for their interesting and useful comments on our manuscript. Our responses to these comments are given below.

Major comment (Section 1): Are the synthetic observations used in this analysis representative of model tuning?

The authors argue that producing one tuned configuration of a climate model underestimates the size of the parameter space (no disagreement there); that the tuning process is often done one parameter at a time, underestimating the non-additive effects of varying multiple parameters at once (only mild disagreement there); and that an analysis based on an ensemble of ERF estimates from tuned models underestimates the true ERF uncertainty. For this last conclusion to be true, the range of ERFs that a single model can plausibly produce ("plausible" meaning consistent with the observed present-day climate) must be non-negligible compared to the intermodel ERF spread, as sketched in Figure 10.

In their analysis, the authors subject the different parameter combinations of their emulated aerosol–climate model to a consistency check against nine observables of the present-day climate. This is the emulator-world equivalent of tuning a single climate model to agree with the present-day climate; the further conclusions made in the manuscript about the uncertainty on the model's ERF estimate thus hinge crucially on whether this consistency with observations is indeed equivalent to model tuning.

I am concerned that the answer may be "no", for the following reason. The observational dataset used here is the July mean conditions over Europe. Europe is a tiny portion of the globe, and July only samples one point in the seasonal cycle of aerosol and cloud properties. Since, as the authors point out, the model uncertainty stems from the interaction of aerosol and host model parameters, it seems like one would want to use the widest spectrum of weather and aerosol conditions available to be able to discard observationally excluded parameter combinations.

As a GCM tuning strategy, a European seasonal approach would fail because the constraint on the global-mean climate would be negligible. I expect that many of the ensemble members that are consistent with the European July observations will have outlandish global-mean TOA fluxes, cloud fields, etc., that would get them rejected in a constraint that uses global observations. This, I expect, would narrow the estimate of the single-model uncertainty compared to what the authors find using their constraint strategy.

The interesting question, in my mind, is how much using global constraints would narrow the uncertainty, i.e., whether Figure 10(b) would still look mostly as it does now or start to look more like Figure 10(a).

I realize that Europe was chosen for a reason, namely that global observations do not exist for all of the fields used as constraints (CCN; decadal trends of surface shortwave radiation and AOD). The most accurate equivalence with GCM tuning would probably be to use global fields where they are available and European fields (but for more than a single month) otherwise.

We don't tackle the problem of global mean ERF or global tuning; the paper is focused solely on Europe. The word tuning doesn't need to imply global. Aerosol models are frequently tuned or adjusted in some way (call it what you will) to agree with observations in a particular region or environment, so it's this process we are drawing comparisons with. We restrict the analysis to Europe because there is a distinct set of parameters that are important to aerosol properties and ERF uncertainty in each region (see Lee et al, 2016: 10.1073/pnas.1507050113 for a map of clusters of uncertain parameters). Because parameter sensitivities vary regionally, to constrain the global forcing you would need to constrain the different regions individually – i.e., in the aerosol world it doesn't make much sense to talk about global tuning. So our European study is like a mini version of what one would need to do globally. If it doesn't work for Europe (with a distinct set of parameters) then it won't work globally either. Using global ToA flux would have limited effect because of many regional compensating factors. We tested this in Regayre et al, 2018 (10.5194/acp-18-9975-2018; section 3.5).

Another reason to use Europe was to relate to the Cherian et al, 2014 study where European surface solar radiation trends were used as an emergent constraint on regional and global forcing. As we state in our conclusions, we think the constraint will be weaker than they calculated because of unconstrainable uncertainties in the individual models.

So to answer the question "how much using global constraints would narrow the uncertainty", the answer is that it would narrow the *global* ERF more than we would currently achieve. But a global constraint won't have much effect on Europe (again, because of the distinct set of parameter uncertainties). We only ever refer to Europe, and we maintain that focus when we compare with the Cherian et al Europe SSR constraint on Europe-mean ERF in Fig 10.

Minor suggestion (Section 2): Some of the conclusions of the paper could be phrased as recommendations on the future direction of the aerosol forcing community

This is an interesting suggestion and we are pleased that the reviewer has identified the potential impacts that our work and results could have in the aerosol forcing community. We did try to draw out the implications. We prefer not to completely change the style of the conclusions because we think it's important that we include a synthesis of findings and integration with current knowledge, rather than recommending future studies. We think recommendations are more appropriate for perspective articles and reviews. In response to some specific suggestions we have tried to highlight the recommendations a bit.

Apart from the main finding of large single-model uncertainty, there are other interesting findings in the paper that I think are underemphasized:

- **The emulator–PPE method is a necessary step toward estimating the model uncertainty correctly. But to narrow the uncertainty range, it will also be necessary to assemble the most powerful (i.e., discriminating) combination of observables. Therefore, I encourage the authors to advocate for both of these necessary steps at once – the more they go hand-in-hand, the more efficient we are likely to be at making progress. (For example, the emulator–PPE method could point to which missing observable would provide the greatest additional constraint if it were added to the existing set of observables; see my comment on Figure 9 in the next section.) This sentiment is kind of there in the last sentence of the abstract and in the text (p. 17, l. 20; p. 23, l. 16), but it would be a missed opportunity not to phrase it more forcefully and more positively in the abstract.**

We have added a sentence on this to the end of abstract to highlight this point. The end of the abstract now reads:

"…However, the uncertainty in the aerosol ERF after observational constraint is large compared to the typical spread of a multi-model ensemble. Our results therefore raise questions about whether the underlying multi-model uncertainty would be larger if similar approaches as adopted here were applied more widely. ***The approach presented in this study could be used to identify the most effective observations for model constraint.*** It is hoped that aerosol ERF uncertainty can be further reduced by introducing process-related constraints, however, any such results will be robust only if the enormous number of potential model variants is explored."

- **The aerosol forcing change over the coming decades is a much easier target than the forcing relative to preindustrial conditions, assuming the emissions changes are known. Perhaps it is time to move away from ERF as the holy grail of the field and instead focus on future aerosol emissions scenarios.**

We agree, and have made this point in at least a couple of previous papers. There is a paragraph (5th of the conclusions in the original manuscript; 6th paragraph in the revised manuscript) dealing with this point, but we agree it could be focused. We now write:

"Observational constraint using nine observations has the potential to reduce the uncertainty in aerosol ERF ***slightly more over a multi-decadal period than over the full industrial period***:"

The rest of the paragraph then explains why this is important. Then we add at the end of the paragraph:

"***A shift of emphasis of the research community towards trying to constrain decadal forcing uncertainty, instead of industrial era forcing, is likely to accelerate progress.***"

- **AOD is a terrible variable if your aim is to understand aerosol–cloud interactions. We should figure out something better. Ed Gryspeerdt's work seems to show that aerosol index is a much better proxy for CCN. If the authors' model diagnoses AI, it would be easy for them to refute or corroborate this result.**

Unfortunately, we cannot compute AI from the model runs of this PPE and so we cannot evaluate this here. However, we disagree that AOD is unambiguously a "terrible variable" for constraining forcing. This point of view is maybe based on the fact that AOD and CCN are not closely related but AI and CCN are more related. However, our study shows that neither CCN nor AOD alone provides a strong constraint on ERF (Fig 7). In that sense you might argue that CCN is also a terrible variable. The key for model constraint and uncertainty reduction is to find *combinations of observables* that are sensitive to the same set of process or emission uncertainties as ERF. So *all* observed variables are helpful in their own way. It is possible that AI will be useful for direct extrapolation back to the pre-industrial period, but it will not necessarily be useful for model constraint. We intend to include AI in the output diagnostics of a future PPE that is in the planning stage, with which we will be able to investigate the potential of AI for model constraint.

We have added a new paragraph in the conclusions (after the 4th paragraph in the original manuscript):

"***It is often argued that AOD is a poor variable to use for understanding aerosol-cloud interactions. However, our results show that even the most strongly related measurement (CCN) also does not***

provide a strong individual constraint on ERF_{ACI} (Figure 7). It is doubtful that other derived variables like aerosol index will be any better. The key to model constraint is to find combinations of observations that help to constrain ERF: Individual constraints are unlikely to be effective, although they may appear to be effective if the model uncertainty is not fully sampled."

These results have a direct bearing on where the aerosol forcing community would best invest its efforts in order to reduce the forcing uncertainty. I understand that the authors want to keep the focus on the main result (the large single-model uncertainty), but they might think about ways to give these other findings a prominent place as well. For example, in our recent review article on radiative forcing by aerosol–cloud interactions (https://doi.org/10.1007/s40641-018-0089-y), we eschewed the traditional re-listing of conclusions at the end of the paper in favor of making recommendations to the forcing community. Mentioning these recommendations in the abstract may be appropriate as well.

We understand how useful a set of recommendations to the aerosol forcing community would be, but we have decided not to fully re-work our conclusions in this paper to a set of recommendations. The presented study, based only on synthetic observations, corresponds to a specific stage in our overall work to understand and constrain the aerosol forcing uncertainty in this aerosol-climate model. Our current and future work is now expanding this work to use real observations in the constraint methodology, and we are learning even more as we move forward with this approach. We feel that recommendations to the community must take into account findings from the complete constraint process in which real observations are used. We therefore elect to defer producing such a set of recommendations at this earlier stage to a future time when we also have a comprehensive evaluation of model constraint with real observations.

Minor Comment 1: Table 1: A bit more detail on parameters 8 and 9 would be nice. (Threshold in what variable? Rate of change with respect to what?)

Full descriptions of the parameters perturbed in this PPE are given in Yoshioka et al, 2018 and Regayre et al, 2018. We therefore do not repeat this full detail in this publication. We have made it clearer to the reader where more detailed descriptions of the parameters can be found by moving the reference for this from the end of paragraph 3 (page 8 line 25 in the original manuscript) to paragraph 2 in Section 2.3. Paragraph 2 of Section 2.3 now reads:

"The 27 perturbed parameters are listed in Table 1. They are categorized as either aerosol (aer) or atmospheric (atm) according to their role in the model. To define the set of parameters we used expert elicitation and carried out one-at-a-time parameter perturbation screening experiments to quantify the effect of individual parameter perturbations away from the default setting. ***The selected parameters are described in more detail in related papers (Regayre et al., 2018; Yoshioka et al., 2018)***"

For the specific parameters mentioned here, we have adapted the text in Table 1 to provide further clarity.

For parameter 8: Dec_Thres_Cld, we have changed the description to: "***Threshold for the ratio of buoyancy consumption to production before decoupling occurs***"

For parameter 9: Fac_Qsat, we have updated the description to: "Rate of change in convective parcel maximum condensate ***with altitude***"

Minor Comment 2: p. 10, l. 8: Do we know that the validation carries over to this new emulator? What is the rationale for not simply using the emulator that is definitely validated?

We are confident that the validation does carry over to the new emulator built using all simulations. The rationale for doing this is to incorporate all of the information that we have from our PPE into the emulator model we use.

We do investigate the validity of the final emulator using a process called 'leave-one-out' validation on the merged set of simulations. In this procedure we remove a simulation from the set, build a new emulator and then use that emulator to predict the removed simulation, repeating this for all simulations in the set in turn. This procedure has returned good validation prediction results in all cases, and Figure A below shows an example of this validation for our present-day AOD emulator. The plot shows that the real model output and corresponding Leave-one-out emulator predictions (with 95% credible intervals) strongly correlate along a line of equality. Hence, this alternate validation indicates that the addition of the extra simulations in the construction of the emulator does not deteriorate the emulator quality and we have confidence in this approach.

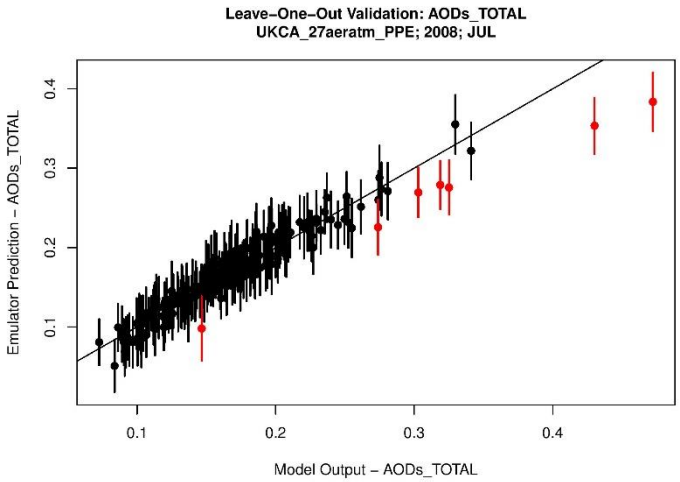

Figure A: Example output of a 'Leave-one-out' validation test of the final emulator.

We have added the following sentence to page 10, line 9 of the original manuscript (also page 10, line 9 of the revised manuscript) to detail that we use a leave-one-out validation procedure to verify the quality of the final emulator:

"***A 'leave-one-out' validation procedure (where each simulation in turn is removed from the merged set, and a new emulator is constructed and used to predict that removed simulation) is applied to additionally verify the quality of our final emulator.***"

Further explanation on our rationale for this is given below for the reviewer's interest.

Given a good validation of an initial emulator based only on the original training runs, we could certainly use this emulator for our analysis and have confidence in the results. However, the validation simulations are purposely placed into gaps between the locations of the training simulations and so including this extra information enables us to better guide/tweak the predicted

emulator response in these gaps to be even more in-line with the underlying 'true' model output. We believe that if a good emulator is achieved based on only the training runs, then a further emulator based on both the training and validation runs must be of similar quality, if not better, given the inclusion of the extra information that these extra runs provide.

Minor Comment 3: p. 10, l. 14: Do the two approaches (elicited vs uniform PDF) give very different results?

In general, we have found that the overall constraint achieved on regional average model outputs (aerosol properties and forcings) is of a very similar nature under both approaches (using the elicited v's using the uniform pdfs). However, there are differences in how the generated large sample of model variants (size 4 million) covers the multi-dimensional parameter uncertainty and hence the proportion of that sample that is retained on application of the constraint methodology.

Using elicited pdfs rather than uniform pdfs when sampling over the multi-dimensional parameter uncertainty space essentially applies a further constraint to that parameter space, sampling much less in areas that the experts believe to be highly unlikely. This means that the sample of model variants tested against the observations in the constraint procedure is more concentrated towards the most likely (as evaluated by our experts) parameter values in the parameter ranges. The areas sampled less tend to be around the edges of the multi-dimensional space.

On constraint we have found that using the elicited pdfs means we retain more of the 4 million model variants and leads us to a slightly stronger constraint on the parameter space itself (e.g. on the marginal parameter space in Figure 9). We reject more variants using uniform pdfs as the corresponding sample under this assumption covers more densely the parts of the parameter space that were deemed unlikely to be plausible by the experts in the elicited pdfs. This result shows some consistency between the model behaviour and the expert judgements, which gives us confidence in using the elicited pdfs. Hence, in the results presented we choose to use the elicited pdfs and so include all available information in the sampling in order to determine the full potential of possible constraint on ERF that is achievable with our approach.

Minor Comment 4: p. 11, l. 10: This does not make intuitive sense to me; if the "truth" is chosen to lie at one the edge of parameter space, shouldn't ensemble members from the opposite direction of that edge be penalized much more strongly than they would be if the "truth" were chosen to lie in the center?

Yes, it is correct that choosing the "truth" to lie in a different part of the parameter space can affect the corresponding 'observed' output values and so change the model variants (and hence parts of parameter space) that are retained on constraint. However, the effect of this change on model output quantities like aerosol forcing is dependent on the way in which the response surfaces of these outputs vary over the multi-dimensional parameter space, and so it is difficult to predict. The compensating effects of the different parameters in the multi-dimensional space can mean that different areas of parameter space lead to similar output values.

For the model run we investigated as a marginal set of synthetic observations (our validation run 27, which had several parameter values towards the edges of their uncertainty range), we found that the achieved overall constraint on each of the forcing variables (using all observations together) was slightly greater and we also retained fewer model variants in the constrained sample. For ERF we

saw a further reduction in the standard deviation of around 12%, however this reduction was much lower for ERF_{ARI} at only 1%. Even so, the relative individual constraint effect of each observational type (Figure 7) was of the same nature/order for all four forcing variables (ToA flux is the most effective for ERF and ERF_{ACI}, the sulphate concentration is most effective for ERF_{ARI} and ERF_{ARIclr}, and in general the achieved constraint from the individual observable aerosol properties is weak). Hence, the overall conclusions of the study did not change on using a different observation set. The resulting constrained set of model variants still corresponds to a large multi-dimensional area of parameter space with a lot of equifinality.

Given these results, and so as not to over-complicate the presented study, we made the decision to focus on just one observation case and the centralised set (with parameters set to their elicited median values) seemed most appropriate for a general example of our approach.

Minor Comment 5: p. 14, l. 2: Much as I hate linear correlation coefficients, perhaps it would be useful to tabulate them to make it easier to follow this discussion. I am having trouble reading them off Figure 2.

We have added the following table of linear correlation coefficients into Section 3.1 of the manuscript:

ToA Flux	0.20	0.00	0.23	-0.30	0.16	0.10	0.04	-0.03	-0.59	-0.65	0.25	-0.05
	ΔSSR	-0.04	0.22	-0.72	0.12	0.49	-0.04	0.11	-0.32	-0.22	-0.52	-0.61
		CCN	0.46	-0.20	0.21	0.03	0.37	0.09	-0.14	-0.14	-0.01	-0.09
			AOD	-0.50	0.88	0.59	0.66	0.59	-0.33	-0.27	-0.33	-0.55
				ΔAOD	-0.32	-0.71	-0.17	-0.20	0.48	0.38	0.55	0.76
					$PM_{2.5}$	0.54	0.69	0.54	-0.21	-0.15	-0.33	-0.44
						Sulphate	0.47	0.64	-0.41	-0.28	-0.67	-0.77
							OC	0.62	-0.34	-0.31	-0.20	-0.30
								BC	-0.27	-0.22	-0.28	-0.46
									ERF	0.98	0.16	0.48
										ERF_{ACI}	-0.04	0.33
											ERF_{ARI}	0.83
												ERF_{ARIclr}

Table 3: Pearson linear correlations (r) between the PPE member regional mean model outputs for Europe in July, for the aerosol properties used as constraints, corresponding to the pairwise scatter plots in Figure 2.

We have edited the first paragraph of Section 3.1 to say:

"Figure 2 shows pairwise scatter plots of the PPE member output (Europe July-mean), which provides an overview of the spread of the model outputs as well as the relationships between the variables. ***We further quantify any linear relationships between the variables using the Pearson correlation coefficient (r) in Table 3.***"

We have adjusted the numbering of all following results tables in the manuscript to incorporate this table into Section 3.1.

Minor Comment 6: Speaking of Figure 2:

- **Nice to run into fellow R users.**
- **All the labels are tiny! par(cex) may help.**

It is difficult with so many small plots in the same figure to make the labels any larger. We have used 'par(cex)' to adjust the size of the axis labels to gain the best possible labelling size whilst avoiding axis values overlapping and becoming unreadable.

- **Point clouds are hard to interpret. Perhaps the relationships between variables would be easier to interpret as color maps of the 2D densities of the emulator results?**

The purpose of this figure is to look at the direct model output from the 191 individual model runs before any interpolation over the parameter uncertainty space (using the emulators) is applied. With only 191 points in each case, we do not have enough data to generate robust density contours and therefore use scatter plots.

- **Raster graphics are an abomination. R supports PDF output; this is the best option in general, and in particular for an information-rich figure like this one, where the reader will want to zoom in on interesting features.**

Figure 2 has been re-made using postscript and PDF graphics, and the new version is much improved for zooming in on interesting features. The new Figure 2 is below (at the end of this comment response)

These comments apply to Figure 3 as well, where additionally I see funny boxes around some of the panels at certain magnifications.

We have looked in detail at Figure 3 and cannot find any issues with the figure on magnification. We therefore have not changed this figure. We will check the proofs.

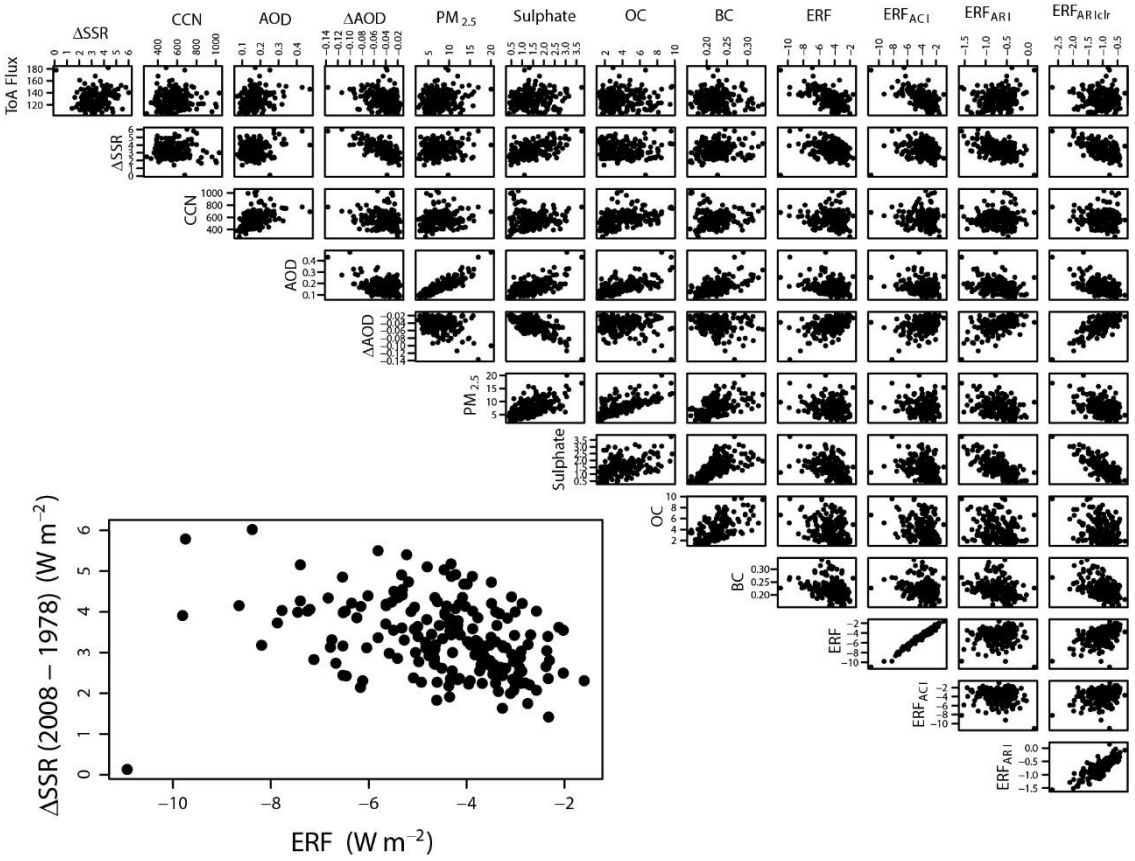

Figure 1. Pairwise scatter plots of the PPE member regional mean model output for Europe in July, for the aerosol properties used as constraints: ToA flux (W m^{-2}), change in SSR (ΔSSR, W m^{-2}) between 1978 and 2008, CCN conc. (cm-3), AOD, surface mass concentrations of **$PM_{2.5}$**, Sulphate, OC, and BC (μg m^{-3}), the changes in AOD (ΔAOD, W m^{-2}) between 1978 and 2008, and the 1850-2008 forcing variables: aerosol ERF, ERF_{ACI}, ERF_{ARI} and ERF_{ARIclr} (W m^{-2}).

Minor Comment 7: Figure 4: In this model, ERF variability is dominated by ERFaci variability (see Figure 2). Yet the model does not care one bit about aerosol–precipitation interactions, at least not wet scavenging. I believe this is quite different from other GCMs. Any idea why? (Not necessarily something that needs to go into the paper.)

The cloud state is sensitive to precipitation scavenging, but the ERF_{ACI} is not. The ERF is obviously a change from pre-industrial to present-day, so it is not obvious that if you change the aerosol/cloud state in both periods that the ERF will remain sensitive to scavenging.

Minor Comment 8: p. 25, l. 15: In GCM tuning, this would routinely be done; see my main point of criticism.

OK.

Minor Comment 9: Figure 9: Does the converse also hold (that observations of decoupling would be a useful constraint on the model)?

We do not know what the reviewer means by “observations of decoupling”, and so we are not sure on what is being referred to here. We are therefore unable to comment.

Minor Comment 10: p. 27, l. 5: But Cherian et al. do this using models tuned to reproduce the global climate; see my main point of criticism.

Yes, we agree that Cherian et al, 2014 achieve their constraint using a set of models that have each been 'tuned' in some way to re-produce the global climate. We compare with our 'tuned' model for Europe (which includes the frequently-tuned quantity, the ToA flux). However, in the selection of each individual tuned model, the influence of parametric uncertainty has not be rigorously explored. This means that there will be many other equally plausible tuned versions of each given model that agree with the observations and could have been selected, as we have shown in this study for the single tuned model HadGEM3-UKCA. We have shown that the predicted forcing from the set of the equally plausible model variants (obtained through comparison to a diverse set of aerosol observations) is wide-ranging. This implies that tuning does not necessarily directly reduce the model spread and implies that emergent constraints such as the one obtained by Cherian et al, 2014 likely underestimate the true spread of forcing.

Minor Comment 11: p. 28, l. 10: AOD multidecadal change appears to be double-counted.

This has been corrected in the manuscript. This sentence in the first paragraph of the conclusions (page 30 lines 5-9 of the revised manuscript) now reads:

"The primary objective of our study was to determine how much uncertainty could remain in an aerosol-climate model when it is constrained to match combinations of observations that define the base state of the model: top-of-atmosphere upward shortwave flux, aerosol optical depth, PM2.5, cloud condensation nuclei, concentrations of sulphate, black carbon and organic material as well as multi-decadal change in surface shortwave radiation and aerosol optical depth."

Minor Comment 12: p. 29, l.1: Would AI or fine-mode fraction work better? I think the authors have the opportunity to make a significant statement here about whether there is a way forward from AOD, which is known to be a poor CCN proxy, via other proxies. See my point in the recommendations section above.

Please see our response to the Minor suggestion (Section 2) above (at the end of page 3). We are not able to evaluate the effects of AI with this PPE, but we doubt that AI would provide any better constraint on the aerosol ERF than AOD. AI might be a better extrapolation variable, but this has not yet been directly demonstrated.

Minor Comment 13: p. 31, l. 6: I don't understand the point about cancellation of correlated errors, but I would like to. Perhaps the authors could elaborate.

We mean that a model might have a large bias in AOD and in CCN (so neither can be simulated well) but if both of these model variables are biased for the same reason (e.g., incorrect aerosol deposition rates), then the ratio of CCN/AOD might be accurately simulated by the model.

Minor Comment 14: Craig 1997 has a bunch of cryptic initials instead of editor names.

This reference has been corrected in the manuscript.

Minor Comment 15: Gryspeerdt 2017 a and b are the same publication.

This has been corrected in the manuscript.

Minor Comment 16: Stier ACPD 2015 has been superseded by Stier 2016 ACP (https://www.atmos-chem-phys.net/16/6595/2016/acp-16-6595-2016.html).

This reference has been corrected in the manuscript.

Minor Comment 17: Penner 2011: the DOI looks strange.

This reference has been corrected in the manuscript.

Minor Comment 18: Pujol 2008: check that this is still up to date with citation("sensitivity").

This reference has been checked in R and corrected in the manuscript.

Minor Comment 19: Zhang 2016: Toshi Takemura's name is misspelled.

This reference has been corrected in the manuscript.

References:

Cherian, R., Quaas, J., Salzmann, M. and Wild, M.: Pollution trends over Europe constrain global aerosol forcing as simulated 25 by climate models, Geophys. Res. Lett., 41(6), 2176–2181, doi:10.1002/2013GL058715, 2014.

Lee, L. A., Reddington, C. L. and Carslaw, K. S.: On the relationship between aerosol model uncertainty and radiative forcing uncertainty., Proc. Natl. Acad. Sci. U. S. A., 113(21), 5820–7, doi:10.1073/pnas.1507050113, 2016.

Regayre, L. A., Johnson, J. S., Yoshioka, M., Pringle, K. J., Sexton, D. M. H., Booth, B. B. B., Lee, L. A., Bellouin, N., and Carslaw, K. S.: Aerosol and physical atmosphere model parameters are both important sources of uncertainty in aerosol ERF, Atmos. Chem. Phys., 18, 9975-10006, doi:10.5194/acp-18-9975-2018, 2018.

Yoshioka, M., Regayre, L., Pringle, K. J., Mann, G. W., Sexton, D. M. H., Johnson, C. E. and Carslaw, K. S.: Perturbed parameter ensembles of the HadGEM-UKCA composition-climate model to explore aerosol and radiative forcing uncertainty, J. Adv. Model Earth Syst., in-prep, 2018.

In our response, reviewer comments are marked in bold, our responses and original text in plain text, and altered text in the paper in bold italic. Additionally we highlight altered text in the tracked changes pdf as requested.

Response to reviewer 2 (Anonymous reviewer)

We thank the reviewer for their interesting and useful comments on our manuscript. Our responses to these comments are given below.

Minor Comment 1: p1 l29 "improvements in the physical realism"... I don't think Mann et al 2014 is the right citation at exactly this point.

We have changed this to "Although extensive improvements in the physical realism of aerosol-climate models have been made in recent years (Ghan and Schwartz, 2007***), resulting in a set of quite sophisticated models (***Mann et al., 2014)." Mann et al, 2014 is really the only reference for where a large set of microphysics models was assembled.

Minor Comment 2: p2 l2: "although the set of models is different to those used to assess aerosol microphysical properties in Mann et al. (2014)," not really an argument for the stubbornness of the ERF uncertainty, can be omitted here.

We have deleted that sentence. It wasn't really an argument, but really just a reminder that we should not compare these two inter-comparisons (Mann et al, 2014 microphysics models and Boucher et al, 2013 climate models) – they are completely different models.

Minor Comment 3: P2 l17 I think this paragraph and equation is misleading in pretending "that the forcing depends on the interlinked sensitivities of aerosols, clouds and their radiative properties to changes in aerosol emissions". Direct radiative effects, fast adjustments are not readily folded in into this equation. Please rephrase.

We are referring only to the aerosol-cloud forcing here. This equation is not pretending anything; it is the community's main approach to understanding how aerosol emission changes affect cloud properties. We have re-written the start of this paragraph (3rd paragraph in the introduction) on page 2 line 14 to clarify this applies only to aerosol-cloud forcing:

"There are three ways in which observations help to constrain the uncertainty in aerosol ERF. The first, ***which applies to the aerosol-cloud-related forcing,*** is based on..."

Minor Comment 4: P3 l23: "there is no equivalent to Equation 1 defining how a bias in simulated aerosol properties affects the forcing " => I think this is overly critical to bias inspections. An underestimate in fine mode AOD or bias in absorption can be translated in forcing bias. Measurements of fine mode AOD estimates can constrain anthropogenic AOD to some extent. And there might be other clever interpretations of bias. Please rephrase.

We disagree. In fact this is one of the main results of our paper: we show that aerosol-radiation interaction forcing (direct effect) is not strongly constrained by state variable measurements (AOD, etc.). There are many ways in which a model can be configured to get a particular AOD, but these

model variants (as we call them) predict very different forcings. To make this clear, and to signpost the result, we add at this point:

"(i.e., there is no equivalent to Equation 1 defining how a bias in simulated aerosol properties affects the forcing). One aim of our study is to make that link***, and we show in section 3.5.1 that observational constraint of many state variables only weakly constrains the direct and indirect radiative forcings.*****"**

Minor Comment 5: P3 l31 "Model variants that produce implausible results are rejected and, likewise, the forcings that they calculate are also rejected. " => would be nice to explain this at this point a bit more. Do you look at all observations at the same time? What is the criterion for rejecting?

The sentence referred to here is in the introduction section where we very briefly summarise / introduce the approach we take in this study. We therefore don't want to go into too much detail, as full details of the methodology and constraint approach are given in the following methodology section. Hence, we have only added very brief extra explanations of these points in this introductory paragraph/section of the manuscript.

In the paragraph on page 3, line 31 of the original manuscript (page 3, line 32 of the revised manuscript), we have added:

"... Model variants that produce implausible results ***(i.e., output outside of an observation's estimated uncertainty range)*** are rejected and, likewise, the forcings that they calculate are also rejected. A similar constraint methodology has been applied to..."

And we have added the following sentence to the end of this paragraph on page 4 line 3 of the original manuscript (page 4, line 4 of the revised manuscript):

"***We constrain using each aerosol/cloud observation individually and combinations of all observations.***"

We have then added more detail on our criteria for retaining/rejecting model variants in the constraint process in Section 2.7 (Identification of plausible model variants) of the methodology section, to make this process clearer. The start of Section 2.7 now reads as:

"Observationally plausible model variants are defined to be those that simulate aerosol and radiation properties within the uncertainty ranges of the observations, defined in Table 2. ***As we use statistical emulators to generate the simulated output values for each model variant, rather than using the climate model directly, an emulator prediction error φ (valued at one standard deviation on the emulator prediction from the Gaussian process uncertainty) is also taken into account. Hence, for a given observed variable, a model variant is rejected as implausible if the range defined by its emulator prediction +/- φ lies outside the corresponding observation's uncertainty range in Table 2. Furthermore, for a joint observational constraint we retain only the model variants that are classed as plausible for all individual observation types that make up the joint constraint.***"

Minor Comment 6: P5 l9 "The analysis is restricted to Europe for the month of July. We do this primarily because regional observations can provide a better constraint on model uncertainty than global mean observations... but with the disadvantage of being less straightforward to understand. . . . We choose Europe because there are many long-term measurements" => I don't buy these arguments. With synthetic observations this should not be a big problem to do globally. There are no long term measurements used. I assume this is done to save computer time. I think its ok to use just Europe and just July. But the discussion should be more honest and open here. Paragraph please rewrite.

There was no intension on our part to not be honest and open in terms of our arguments for only using observations over the Europe region in July for this study.

Our full reasoning to base the presented study on only Europe in July is as follows:

- Previous work (Regayre et al, 2018, for example) has shown that using global mean quantities for constraint can mask many compensating regional parameter effects, leading to a very weak 'watered down' constraint on both the parameter space and model outputs like forcing that can be difficult to interpret.
- To constrain forcing globally we need to constrain the parameters that affect the forcing across the globe. Regayre et al, 2018 show that different parameter sources control the uncertainty in aerosols and forcing in different regions. Therefore, the global problem essentially breaks down to be the sum of constraining the forcing in a set of key regions, of which Europe is one. The Europe region in July provides a single region/month for which a distinct set of parameter uncertainties affect ERF. If we cannot constrain the forcing regionally in Europe, then we are unlikely to obtain a constraint on a global/multi-region scale. Hence, Europe in July provides us with the full insight we aim for here on the potential of our approach.
- It is true that our analysis is highly computational and generates a significant amount of data. We have investigated other regions in this work, including China and the North Pacific (not shown), but including more regions in the presented study would only significantly expand the results in terms of quantity and complexity, with no real gain as to our actual aim of establishing the overall potential for constraint with our methodology.
- Real observations of multiple aerosol observable quantities are sparse in many regions around the globe, and temporally (Reddington *et al.*, 2017), but Europe is a region for which a diverse set of aerosol observations are available. These observations provide realistic estimates of observational uncertainty for the synthetic study.
- The presented study was a specific stage in our model evaluation work, at which we aimed to test our methodology for constraint using synthetic observations before moving forward to our now current work where we are looking at using real observations. Using Europe in July for our synthetic study supplies a good test for evaluating the potential constraint we may achieve from using the real observations in the future – a test that we are now working towards verifying.

We have edited the penultimate paragraph in the introduction section at page 5 line 9 of the original manuscript (page 5 line 11 of the revised manuscript) to better reflect these reasons (The start of this paragraph also contains revisions with respect to our response to Reviewer 3's minor comment 2):

"The analysis is restricted to the region of Europe (defined in this study by the longitude range: 12°W to 41°E, and latitude range: 37.5°N to 71.5°N) for the month of July. We take a regional approach primarily because regional observations provide a better constraint on model uncertainty than global mean observations (Regayre et al., 2018). The sources of uncertainty in aerosols and forcing vary regionally (Lee et al., 2016; Reddington et al., 2017; Regayre et al., 2015)***. Therefore,*** a global analysis would essentially be a scaled-up version of what we present here – i.e., a set of ***regional evaluations.*** We choose Europe ***in July as this is a region and month for which a distinct set of parameter uncertainties affect the aerosol properties and the ERF, providing a good test case for our methodology. Europe is also a region for which a diverse set of*** long-term measurements of different aerosol and radiative properties ***are*** available, ***that*** we can use to inform our assessments of the observational uncertainty."

Minor Comment 7: Chapter 2.1 and 2.2. and 2.3: I think they can be reversed. Some simple questions are not clear to me: Are the simulations global? Is it a one year simulation with a 4 month spinup (eg Sep-Dec of the preceding year) and is then just July analysed? Is the emulator producing global fields, from which data are sampled at European stations?

We have considered the reviewers suggestion to change the order of the sub-sections in our methodology section (Section2) of the manuscript. However, we think that the current order is most suitable, as we prefer to keep the overall summary of our approach at the start (section 2.1), with the different aspects further explained in the order they are mentioned in that summary.

We have clarified the model set up in the final paragraph of section 2.3 (page 9 lines 11-12 of the original manuscript; page 9 lines 15-17 in the revised manuscript):

"... In total 217 perturbed parameter simulations ***of the global model*** were run ***for a full year*** for each anthropogenic emission period (1850, 1978 and 2008 emissions). ***Each simulation had a spin-up period of seven months from a consistent starting simulation, where the parameters were set to their median values for the first four months and the perturbations then applied in the final three months***."

The emulators do not produce global fields. We have clarified this at the beginning of Section 2.4 (page 10, line 6 of the original manuscript; page 10 line 5 of the revised manuscript):

"For each model output (such as ***the regional mean ToA flux, CCN conc., etc. for Europe in July***) we construct a statistical emulator model over the 27-dimensional parameter uncertainty..."

Minor Comment 8: Page 5 counts 191 simulations, while page 9 counts "in total 217 perturbed parameter simulations". Better to harmonize numbers.

The reasoning for this difference is explained in the next sentence on page 9 (lines 12-13). We only use ensemble members that completed the full year of simulation in all periods, which reduces the number of runs used for analysis to 191 from the total of 217 that were originally run. We feel it is important that we are transparent about this and so we continue to state both numbers. We have amended the sentence at page 9 line 12 of the original manuscript (page 9 line 18 of the revised manuscript) to improve the clarity on this point:

"Twenty-five simulations did not complete ***the full*** annual cycle ***so were not used in our analysis. Consequently,*** the ensemble of simulations ***used for analysis*** for each period was made up of the remaining 191 simulations, all of which were used to build the final emulators."

Conclusions: I wonder how general the findings are if the ERF is in essence tested only over Europe and July with synthetic observations, but that might be shown in future publications.

The aim of this paper is to demonstrate the potential for model constraint using multiple observations. As we argue in the paper (and in our replies to the reviewer comments) a global analysis would essentially be a scaled up version of what we are doing here – i.e., constraint of global ERF will be dependent on the extent to which we can constrain the model in all the key regions. Global forcing is the sum of regional forcings, and each region has its own unique combination of uncertainties.

References:

Boucher, O., Randall, D., Artaxo, P., Bretherton, C., Feingold, G., Forster, P., Kerminen, V.-M., Kondo, Y., Liao, H., Lohmann, U., Rasch, P., Satheesh, S. K., Sherwood, S., Stevens, B. and Zhang, X. Y.: Clouds and Aerosols, in Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by V. B. and P. M. M. Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. 571., 2013.

Mann, G. W., et al.: Intercomparison and evaluation of global aerosol microphysical properties among AeroCom models of a range of complexity, Atmos. Chem. Phys., 14(9), 4679–4713, doi:10.5194/acp-14-4679-2014, 2014.

Regayre, L. A., Johnson, J. S., Yoshioka, M., Pringle, K. J., Sexton, D. M. H., Booth, B. B. B., Lee, L. A., Bellouin, N., and Carslaw, K. S.: Aerosol and physical atmosphere model parameters are both important sources of uncertainty in aerosol ERF, Atmos. Chem. Phys., 18, 9975-10006, doi:10.5194/acp-18-9975-2018, 2018.

Reddington, C. L., et al.: The Global Aerosol Synthesis and Science Project (GASSP): Measurements and Modeling to Reduce Uncertainty, Bull. Am. Meteorol. Soc., 98(9), 1857–1877, doi:10.1175/BAMS-D-15-00317.1, 2017.

In our response, reviewer comments are marked in bold, our responses and original text in plain text, and altered text in the paper in bold italic. Additionally we highlight altered text in the tracked changes pdf as requested.

Response to reviewer 3 (Anonymous reviewer)

We thank the reviewer for their interesting and useful comments on our manuscript. Our responses to these comments are given below.

Main comment: One comment I have, is that it should be made more clear in abstract and conclusion, and also some figures and tables, that they use synthetic observations and not real observations. And define synthetic observations the first time it is mentioned.

We have adapted the text in the abstract and conclusions, and in the captions of Table 2 and Figures 3 and 7 to make this clearer throughout the manuscript.

- In the abstract, we have revised the following sentence on page 1 line 18:
  "The model uncertainty is calculated by using a perturbed parameter ensemble that samples twenty-seven uncertainties in both the aerosol model and the physical climate model***, and we use synthetic observations generated from the model itself to determine the potential of each observational type to constrain this uncertainty.***"

- The caption of Table 2 has been adjusted to:
  "**Table 2.** Observed quantities and corresponding uncertainty ranges used for the constraints applied over Europe. Values are a European July mean***, synthetically generated from the model output of a selected PPE member.***"

- The caption of Figure 3 has been adjusted to:
  "**Figure 3.** Calculated uncertainty in the aerosol quantities and aerosol ERF terms from the 4 million member sample. Results are for July-mean over Europe. The red bar shows the assumed range of each ***synthetic*** observation used to constrain the uncertain parameter space and the aerosol forcing uncertainty from Table 2."

- The caption of Figure 7 has been adjusted to:
  "**Figure 2.** The relative constraint achieved for aerosol ERF, ERF_{ACI}, ERF_{ARI} and ERF_{ARIclr} over Europe given the individual ***synthetic*** constraints applied (colours) and the simultaneous constraint (ALL). The relative constraint is evaluated as the ratio of the standard deviation of the forcing in the constrained sample ($\sigma_{constrained}$) to the standard deviation of the forcing in the original, unconstrained sample (σ_{full})."

- In the conclusions section we have added the following sentence to page 28 line 17 of the original manuscript (page 30, line 16 in the revised manuscript).
  "***Using synthetic observations (taken from the output of one of our simulations) we determine the extent of the potential constraint that these nine aerosol and cloud-related properties can generate.***"

Finally, we have defined the term "synthetic observations" at the point it is first mentioned in the body of the manuscript, in the introduction section on page 5 line 6 of the original manuscript (page 5 line 8 in the revised manuscript). The revised text is as follows:

"Although large observational datasets of aerosol in-situ microphysical and chemical properties are available (Reddington et al., 2017), we use synthetic observations here ***– i.e., observations that are generated from a model simulation –*** to postpone addressing some of the challenges faced when comparing model output and in-situ observations (Schutgens et al., 2016a, 2016b)."

Minor Comment 1: Page 7 line 10. Specify that it is biomass burning emissions.

Page 7, line 10 of the original manuscript is an empty line break between paragraphs. However, we think this is referring to page 7 lines 19-20 (page 7 line 23 of the revised manuscript), and have updated the text here as follows:

"Carbonaceous ***biomass burning*** aerosol emissions for recent decades were prescribed using a ten year average of 2002 to 2011 monthly mean data"

Minor Comment 2: Table 2: Indicate that this is not real observations. Useful to define Europe also. In addition to the synthetic observations, real observations are used for ToA flux, am I right?

We have adjusted the caption for Table 2 to be clear that the observations are not real observations. (See bullet point 3 in our reply to the Main Comment above.)

We have updated the text at the end of the introduction section (page 5, line 9 of the original manuscript; page 5 line 11 of the revised manuscript) to more clearly define the Europe region that we have used. Revised text:

"The analysis is restricted to ***the region of*** Europe ***(defined in this study by the longitude range: 12°W to 41°E, and latitude range: 37.5°N to 71.5°N)*** for the month of July."

All observations used in this study, including the ToA Flux observation, are synthetic and come from a model run with all parameters set to their median value from the parameter's distribution that was obtained through our expert elicitation exercise. However, information from real observations (where available) was used to determine appropriate uncertainty ranges on the synthetic observations. For ToA flux, the uncertainty range was estimated to be in line with information from the Clouds and the Earth's Radiant Energy System (CERES) and IPCC uncertainty estimates (Hartmann et al., 2013). The information on the origin of the ToA flux observation used in this study in paragraph 2 of Section 2.6 was incorrect in our original manuscript, and we have amended paragraphs 2 and 3 of Section 2.6 to address this. The revised paragraphs are as follows:

"We use synthetic observations (Table 2) of European July-mean cloud condensation nuclei (CCN) concentration at 0.2% supersaturation at approximate cloud-base height, surface concentrations of PM2.5, mass concentrations of sulphate, OC and BC at the surface, ***the outgoing shortwave radiative flux at the top of the atmosphere (ToA flux),*** AOD at a wavelength of 550 nm, and the change in AOD (ΔAOD) and surface solar radiation (ΔSSR) between 1978 and 2008. The period 1978 to 2008 was originally chosen because it is an interesting period for global and regional forcing changes. Although AOD measurements are not available back to 1978, this is not vital to the present study which aims to assess potential constraint over a period with substantial aerosol changes.

The observation uncertainties are based on our judgement about the combined effect of instrument uncertainties and the uncertainty associated with measurement representativeness (colocation of

high-frequency point measurements within low-spatial-resolution, monthly-mean model output subject to meteorological variability (Reddington et al., 2017; Schutgens et al., 2016a, 2016b). ***Where available, we have used sets of real observations to inform these judgements and estimates. For example, we selected our uncertainty range on the ToA Flux such that it is in line with information from the Clouds and the Earth's Radiant Energy System (CERES) and IPCC uncertainty estimates (Hartmann et al., 2013).*** In the constraint process we also account for the emulator error (i.e., the estimated uncertainty in each of the 4 million points associated with using the emulator instead of the model itself)."

Minor Comment 3: Figure 9: Does the color shading mean anything? Include a colorbar or remove the shading.

The colour represents the marginal normalised sampling density (normalised across parameters) of each input parameter over its range. The parts of the marginal parameter space that are effectively ruled out by the constraint are shown in white (normalised sampling density <0.02).

The plot has been updated and a colour bar has been added. We have also updated the caption of the figure to clearly state the meaning of the colour-scale. The new figure is below.

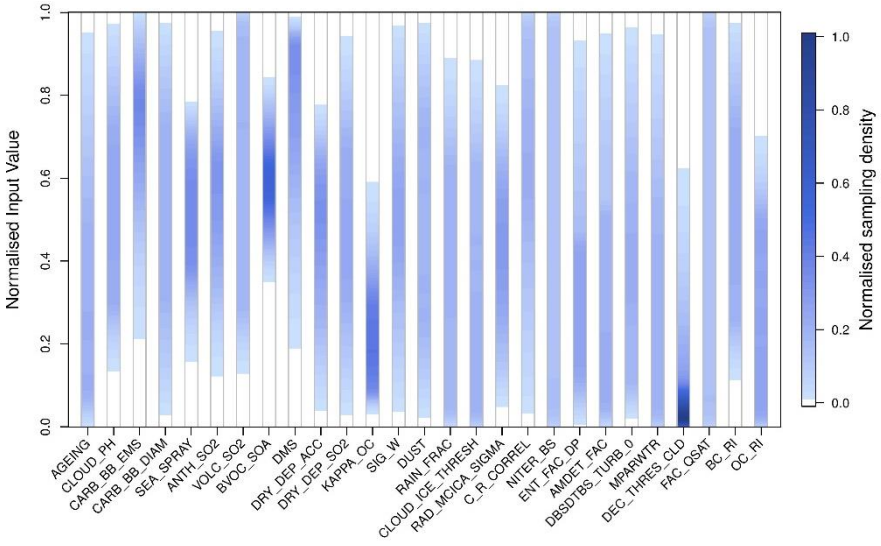

Figure 3. One-dimensional projection of the remaining parameter space after simultaneous constraint of all atmospheric quantities and decadal trends. ***The colour-scale shows the marginal normalised sampling density (normalised across parameters) of each input parameter over its range. Parts of the marginal parameter space that are effectively ruled out are shown in white (normalised sampling density <0.02)***.

We have also updated a sentence on page 24 line 28 of the original manuscript (page 26 line 28 of the revised manuscript) for clarity. This sentence now reads:

"Figure 9 identifies parts of the ***marginal*** parameter space that are ***effectively*** ruled out ***in white***."

The importance of comprehensive parameter sampling and multiple observations for robust constraint of aerosol radiative forcing

Jill S. Johnson[1], Leighton A. Regayre[1], Masaru Yoshioka[1], Kirsty J. Pringle[1], Lindsay A. Lee[1], David M. H. Sexton[2], John W. Rostron[2], Ben B. B. Booth[2] and Kenneth S. Carslaw[1]

[1]School of Earth and Environment, University of Leeds
[2]Met Office, Exeter, UK

Correspondence to: Jill S. Johnson (j.s.johnson@leeds.ac.uk)

Abstract. Observational constraint of simulated aerosol and cloud properties is an essential part of building trustworthy climate models for calculating aerosol radiative forcing. Models are usually tuned to achieve good agreement with observations, but tuning produces just one of many potential variants of a model, so the model uncertainty cannot be determined. Here we estimate the uncertainty in aerosol effective radiative forcing (ERF) in a tuned climate model by constraining 4 million variants of the HadGEM3-UKCA aerosol-climate model to match nine common observations (top-of-atmosphere shortwave flux, aerosol optical depth, $PM_{2.5}$, cloud condensation nuclei, concentrations of sulphate, black carbon and organic carbon, as well as decadal trends in aerosol optical depth and surface shortwave radiation.) The model uncertainty is calculated by using a perturbed parameter ensemble that samples twenty-seven uncertainties in both the aerosol model and the physical climate model, and we use synthetic observations generated from the model itself to determine the potential of each observational type to constrain this uncertainty. Focusing over Europe, we show that the aerosol ERF uncertainty can be reduced by about 30% by constraining it to the nine observations, demonstrating that producing climate models with an observationally plausible "base state" can contribute to narrowing the uncertainty in aerosol ERF. However, the uncertainty in the aerosol ERF after observational constraint is large compared to the typical spread of a multi-model ensemble. Our results therefore raise questions about whether the underlying multi-model uncertainty would be larger if similar approaches as adopted here were applied more widely. The approach presented in this study could be used to identify the most effective observations for model constraint. It is hoped that aerosol ERF uncertainty can be further reduced by introducing process-related constraints, however, any such results will be robust only if the enormous number of potential model variants is explored.

1 Introduction

It has proven extremely challenging to reduce the large uncertainty in aerosol model simulations and the calculated aerosol radiative forcing since pre-industrial times. Although extensive improvements in the physical realism of aerosol-climate models have been made in recent years (Ghan and Schwartz, 2007), resulting in a set of quite sophisticated models (Mann et al., 2014), aerosol model simulations are still surprisingly uncertain – up to a factor ten or more spread in key aerosol properties

among models (Mann et al., 2014). Calculated aerosol radiative forcing also remains stubbornly uncertain among multiple models (Boucher et al., 2013; Myhre et al., 2013), making it difficult to establish the causes of forcing uncertainty. Until the uncertainty is reduced, climate models will not be robust in their predictions of decadal-scale climate change and its global and regional impacts (Andreae et al., 2005; Myhre et al., 2013; Seinfeld et al., 2016).

The uncertainty in aerosol radiative forcing has persisted through multiple generations of climate models because it results from the combined effects of dozens of complex and uncertain climate model processes related to aerosols, clouds, radiation and dynamics. Changes in aerosols cause the entire aerosol-cloud-radiation-dynamics system to respond, resulting in an Effective Radiative Forcing, or ERF (Boucher et al., 2013). The complexity of the processes causing the aerosol ERF (and the fact that it cannot be measured directly) means that it may essentially be treated as a tuneable model quantity (Hourdin et al., 2016; Mauritsen et al., 2012) rather than being properly constrained by extensive measurements that define the state and behaviour of aerosols and clouds. This is not a firm basis for climate projections.

There are three ways in which observations help to constrain the uncertainty in aerosol ERF. The first, which applies to the aerosol-cloud-related forcing, is based on the recognition that the forcing depends on the interlinked sensitivities of aerosols, clouds and their radiative properties to changes in aerosol emissions. For example, the magnitude of the aerosol-cloud interaction component of radiative forcing (R) can be broken down into a product of sensitivities relating the forcing to aerosol emissions (E), cloud condensation nuclei concentrations (N_{CCN}) and droplet concentrations (N_d) (Ghan et al., 2016):

$$\frac{d\ln R}{d\ln E} = \frac{d\ln N_{CCN}}{d\ln E} \times \frac{d\ln N_d}{d\ln N_{CCN}} \times \frac{d\ln R}{d\ln N_d} \quad (1)$$

Relationships between various aerosol, cloud and radiation variables are widely used or proposed as a way of constraining the uncertainty in aerosol-cloud forcing in climate models (Ban-weiss et al., 2014; Grandey et al., 2013; Gryspeerdt et al., 2016, 2017; Gryspeerdt and Stier, 2012; Lebo and Feingold, 2014; McCoy et al., 2016; Quaas et al., 2009, 2010; Terai et al., 2015; Yi et al., 2012; Zhang et al., 2016).

The second aspect of model constraint is to test the model's ability to reproduce observed trends in aerosols, clouds and radiation (Allen et al., 2013; Cherian et al., 2014; Leibensperger et al., 2012; Li et al., 2013; Liepert and Tegen, 2002; Shindell et al., 2013; Turnock et al., 2015; Zhang et al., 2017). For example, Cherian et al. (2014) showed that among several climate models there is a relationship between the simulated trend in European surface solar radiation (SSR) over recent decades and the pre-industrial to present-day aerosol ERF (models with large SSR trends tend to simulate larger ERFs). Cherian et al.

(2014) used this relationship to define the observationally constrained ERF based on the models that simulate SSR trends closest to observations (a so-called emergent constraint).

The third aspect of model constraint is to observationally constrain the model "base state" – i.e., properties like aerosol optical depth (AOD) or aerosol concentrations in a particular period. Considerable effort is put into constraining the model base state because observations are readily available and models that cannot simulate aerosol and cloud properties close to observations would not be trusted to predict changes in these properties over time (which determines the forcing). Models can also be constrained under a range of cloud regimes as well as under pristine and polluted conditions, which will have a bearing on a model's ability to simulate the change from the pre-industrial period to the present-day (Carslaw et al., 2013, 2017). Model skill in simulating AOD was used in the Atmospheric Chemistry-Climate Model Intercomparison Project to screen the models (Shindell et al., 2013) and global AOD reanalysis products have been used to help constrain the aerosol forcing (Bellouin et al., 2013). It is also argued that the wealth of available measurements will help to constrain direct radiative forcing (Kahn, 2012)

There are limitations with all three methods outlined above in terms of constraining the uncertainty in aerosol forcing over periods outside the observational record. The main limitation with the first method (aerosol-cloud-radiation relations) is that there is no guarantee that present-day (or "instantaneous") relationships can be extrapolated to pristine pre-industrial conditions (Penner et al., 2011). Even the most sophisticated approaches still rely on model estimates of how aerosols changed over the industrial period (Gryspeerdt et al., 2017). The same main limitation applies to the second method (aerosol and radiation trends): most data records are quite short so typically do not include pristine pre-industrial-like conditions (Carslaw et al., 2017; Hamilton et al., 2014). With the third method (constraining the state of aerosols, clouds and their radiative properties) it is not obvious how the model accuracy can be related to the uncertainty in simulated radiative forcing (i.e., there is no equivalent to Equation 1 defining how a bias in simulated aerosol properties affects the forcing). One aim of our study is to make that link, and we show in section 3.5.1 that observational constraint of many state variables only weakly constrains the direct and indirect radiative forcings.

In this paper we focus on observationally constraining uncertainty in the base state of an aerosol-climate model as well as trends in radiative properties. Our approach is shown schematically in Figure 1. We begin with a large set of model variants produced by adjusting multiple uncertain model input parameters (a tiny fraction of which would be explored in model tuning). These model variants (parameter combinations) define the prior model uncertainty (which can be defined by a pdf), which we then constrain by identifying variants that produce plausible outputs compared to aerosol and cloud observations. Model variants that produce implausible results (i.e., output outside of an observation's estimated uncertainty range) are rejected and, likewise, the forcings that they calculate are also rejected. A similar constraint methodology has been applied to environmental models (Salter and Williamson, 2016), hydrological models (Liu and Gupta, 2007), galaxy formation (Rodrigues et al., 2017),

disease transmission (Andrianakis et al., 2017), climate models (Murphy et al., 2004; Regayre et al., 2018; Sexton et al., 2011; Williamson et al., 2013)and aerosol models (Lee et al., 2011, 2013; Reddington et al., 2017; Regayre et al., 2018, 2014, 2015). In this paper the observations comprise aerosol and cloud state variables and trends, but the approach could readily be extended to include any observations, such as of aerosol-cloud-radiation relationships. We constrain using each aerosol/cloud observation individually and combinations of all observations.

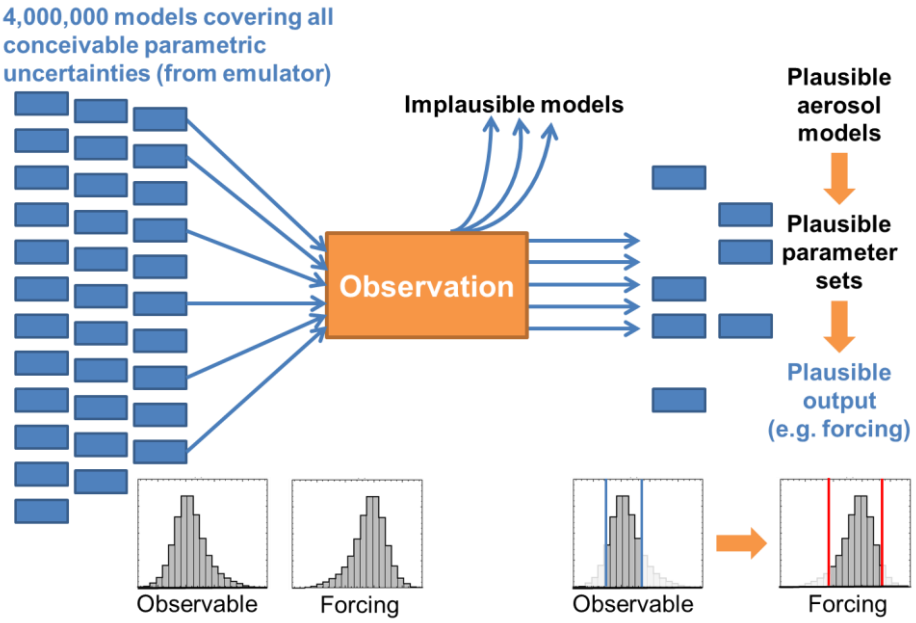

Figure 1. Schematic of the methodology for observational constraint of parametric model uncertainty.

We define observational constraint as *finding the full set of model variants that can be considered plausible against observations*, and from which we can estimate the prior (unconstrained) and remaining (observationally constrained) uncertainty. This approach is different to traditional model tuning, which produces only one result on the right side of Figure 1 with no information about uncertainty. We note, however, that such model adjustments towards observations are often misleadingly called constraint.

The vast majority of observational constraint efforts are severely limited by the very small number of models used, which makes it impossible to reach robust statistical conclusions about model uncertainty. In a multi-model ensemble the number of models is often about ten or so, and in model tuning perhaps only a few dozen parts of parameter space are explored. To get around this problem we build emulators that enable model outputs to be generated for millions of model parameter

combinations (Lee et al., 2011, 2013), which enables us to relate the uncertainty on the left side of Figure 1 (in the form of a pdf) to the observationally constrained uncertainty on the right side.

The main aims of this paper are first to determine how much uncertainty could potentially remain in an aerosol-chemistry-climate model that is tuned to match various sets of observations, and second, how this uncertainty might affect conclusions drawn from multi-model ensemble studies which do not explicitly account for this source of uncertainty. Although large observational datasets of aerosol in-situ microphysical and chemical properties are available (Reddington et al., 2017), we use synthetic observations here – i.e., observations that are generated from a model simulation – to postpone addressing some of the challenges faced when comparing model output and in-situ observations (Schutgens et al., 2016a, 2016b).

The analysis is restricted to the region of Europe (defined in this study by the longitude range: 12°W to 41°E, and latitude range: 37.5°N to 71.5°N) for the month of July. We take a regional approach primarily because regional observations provide a better constraint on model uncertainty than global mean observations (Regayre et al., 2018). The sources of uncertainty in aerosols and forcing vary regionally (Lee et al., 2016; Reddington et al., 2017; Regayre et al., 2015). Therefore, a global analysis would essentially be a scaled-up version of what we present here – i.e., a set of regional evaluations. We choose Europe in July as this is a region and month for which a distinct set of parameter uncertainties affect the aerosol properties and the ERF, providing a good test case for our methodology. Europe is also a region for which a diverse set of long-term measurements of different aerosol and radiative properties are available, that we can use to inform our assessments of observational uncertainty.

The following section describes the aerosol-climate model, the set-up of the simulations and the statistical methodology we use to sample the model uncertainty. Section 3 describes our results, starting with an analysis of the magnitude and causes of model uncertainty. We then examine the effects of observational constraint on the simulated aerosol properties, the multi-century (1850-2008) and multi-decade (1978-2008) aerosol ERF uncertainty and the plausible parameter ranges. In section 4 we estimate the potential implications of our results for multi-model emergent constraint studies and other studies that use observations to screen out models.

2 Methods

2.1 Summary of the constraint methodology

The steps involved are (Figure 1):

1. A perturbed parameter ensemble (PPE) of the HadGEM3-UKCA aerosol-chemistry-climate model (section 2.2) is created to efficiently sample combinations of 27 uncertainties related to the aerosol model and physical processes in the host climate

model (mostly related to clouds). The PPE (section 2.3) consists of three sets of 191 single-year simulations which differ only in the anthropogenic aerosol emissions prescribed (1850, 1978 and 2008). The use of HadGEM enables us to diagnose the aerosol ERF rather than just the cloud albedo forcing as in our previous studies (Carslaw et al., 2013; Regayre et al., 2014, 2015).

2. Emulators are built based on the PPE training data (step 1) which define (within quantifiable uncertainty) how aerosol properties and aerosol radiative forcing vary over the 27-dimensional parameter space (section 2.4). We validate each emulator's ability to reproduce model output, then use them to sample the 4 million Monte Carlo points from the parameter space to produce the set of model variants on the left side of Figure 1. This step is essential because, with 27 dimensions of model uncertainty, the 191 PPE simulations are sparsely distributed. A denser sample of the multi-dimensional parameter space from the emulator enables us to conduct robust statistical analyses.

3. The causes of uncertainty in the aerosol and forcing variables are determined using variance-based sensitivity analysis (section 2.5). This step is not essential for constraining the model, but is useful for understanding which processes in the model account for the uncertainties in the outputs (Carslaw et al., 2013; Lee et al., 2013; Regayre et al., 2018, 2014, 2015).

4. A set of 'synthetic' observations (section 2.6) is created with realistic uncertainty ranges. We use one PPE member to define these synthetic observations.

5. We identify which of the 4 million model variants are consistent with the observations within their individual uncertainty ranges (section 2.7). This reduced set of variants defines the ways in which parameter values can be combined to reproduce multiple observations and is equivalent to identifying several thousand equally plausible tuned HadGEM3-UKCA models. This procedure is often called 'history matching' or 'pre-calibration' (Craig et al., 1997; Edwards et al., 2011; Williamson et al., 2013; Lee et al., 2016; Andrianakis et al., 2017).

6. The reduction in aerosol ERF uncertainty is quantified using the observationally constrained parameter space (section 2.8).

The observational constraint approach we apply here is quite different to aerosol data assimilation (Bellouin et al., 2013), which cannot directly estimate aerosol ERFs nor the uncertainty. In principle, both approaches should generate similar distributions of AOD (the usual assimilated observation variable) if similar observations are used. However, we can directly determine aerosol ERF and its uncertainty by running the plausible model variants in both the present-day (where the model uncertainty was constrained) and the pre-industrial. In contrast, estimation of the ERF using the assimilation approach relies on assumptions about how present-day natural AOD represents pre-industrial aerosols because the approach generates only a present-day aerosol state and not a model that can be used to simulate pre-industrial conditions.

2.2 The HadGEM3-UKCA climate model

We use the UK Hadley Centre Unified Model HadGEM3 (HadGEM3, 2017) incorporating version 8.4 of the UK Chemistry and Aerosol (UKCA) model. UKCA simulates trace gas chemistry and the evolution of the aerosol particle size distribution and chemical composition using the GLObal Model of Aerosol Processes (GLOMAP-mode; (Mann et al., 2010)) and a whole-atmosphere chemistry scheme (O'Connor et al., 2014). The model has a horizontal resolution of 1.25x1.875 degrees and 85 vertical hybrid pressure levels.

The aerosol size distribution is defined by seven log-normal modes: one soluble nucleation mode as well as soluble and insoluble Aitken, accumulation and coarse modes. The aerosol chemical components are sulphate, sea salt, black carbon (BC), organic carbon (OC) and dust. Secondary organic aerosol (SOA) material is produced from the first stage oxidation products of biogenic monoterpenes under the assumption of zero vapour pressure. SOA is combined with primary particulate organic matter after kinetic condensation.

GLOMAP simulates new particle formation, coagulation, gas-to-particle transfer, cloud processing and deposition of gases and aerosols. The activation of aerosols into cloud droplets is calculated using globally prescribed distributions of sub-grid vertical velocities (West et al., 2014) and the removal of cloud droplets by autoconversion to rain is calculated by the host model. Aerosols are also removed by impaction scavenging of falling raindrops according to the parametrisation of clouds and precipitation collocation (Boutle et al., 2014; Lebsock et al., 2013). Aerosol water uptake efficiency is determined by kappa-Kohler theory (Petters and Kreidenweis, 2007) using composition-dependent hygroscopicity factors.

Anthropogenic emission scenarios prepared for the Atmospheric Chemistry and Climate Model Inter-comparison Project (ACCMIP) and prescribed in some of the CMIP Phase 5 experiments are used here. Carbonaceous biomass burning aerosol emissions for recent decades were prescribed using a ten year average of 2002 to 2011 monthly mean data from the Global Fire and Emissions Database (GFED3; (van der Werf et al., 2010)) and according to Lamarque et al. (2010) for 1850. The prescribed volcanic SO_2 emissions combine emissions from the Andres and Kasgnoc (1998) dataset for continuously erupting and sporadically erupting volcanoes and the Halmer et al. (2002) dataset for explosive volcanoes.

Horizontal winds in the simulations are nudged towards European Centre for Medium-Range Weather Forecasts (ECMWF) ERA-Interim reanalyses for 2006 between approximately 2.15 and 80 km using a 6-hour relaxation timescale. Nudging means that pairs of simulations have near-identical synoptic-scale features, which enables the effects of perturbations to aerosol and chemical processes within the boundary layer to be quantified using single-year simulations. Without nudging, the model fields would need to be averaged over several decades in order to produce signals stronger than the noise caused by internal variability

(Kooperman et al., 2012). By nudging horizontal winds but not temperature, liquid water path and atmospheric humidity can respond to aerosol-induced changes in temperature, allowing more of the rapid responses of clouds and radiation to aerosol perturbations to be captured.

Each simulation was subject to a four month spin-up period with parameters set to their median values. Parameter perturbations were then applied distinctly to individual ensemble members and spun-up for a further 9 months. We analyse the data from July for each simulation following the spin-up period. The calculation of the aerosol ERF and its components is described in section 2.8.

2.3 Perturbed parameter ensemble

A perturbed parameter ensemble (PPE) is a set of simulations with excellent space-filling properties that provides information about model output across the multi-dimensional space of uncertain model input parameters. The PPE, described in detail in Yoshioka et al. (2018), was specifically designed to sample aerosol as well as host physical climate model parameters of importance to the aerosol ERF. Regayre et al. (2018) show that host model parameters cause most of the uncertainty in the radiative state of the atmosphere but aerosol parameters contribute more to the uncertainty in the change-of-state uncertainty (aerosol ERF) .

The 27 perturbed parameters are listed in Table 1. They are categorized as either aerosol (aer) or atmospheric (atm) according to their role in the model. To define the set of parameters we used expert elicitation and carried out one-at-a-time parameter perturbation screening experiments to quantify the effect of individual parameter perturbations away from the default setting. The selected parameters are described in more detail in related papers (Regayre et al., 2018; Yoshioka et al., 2018).

Eighteen parameters related to aerosol and precursor gas emissions, deposition and aerosol processes were perturbed based on their importance as causes of uncertainty in aerosols and aerosol-cloud forcing (Lee et al., 2013, Carslaw et al., 2013, Regayre et al., 2014, 2015). Several parameters available in the HadGEM3-UKCA model but not in the chemistry transport model were included after analysing the one-at-a-time perturbation screening experiments. These are the updraft velocity in shallow clouds, the fraction of large-scale cloud in which rain-scavenging of aerosols can occur, and the refractive indices of BC and OC. In some cases we perturbed similar parameters as in Regayre et al. (2014) but these parameters are handled differently within the HadGEM model. These are the dry deposition velocity of SO_2, dust emissions, and the fraction of ice in mixed-phase clouds above which aerosol scavenging is suppressed.

Nine physical model parameters were perturbed. These were selected from a much larger set tested by the UK Met Office in developing their ensemble prediction system (Sexton et al., 2017) based on their potential to contribute to uncertainty in a

broad range of aerosol, cloud and radiation properties; in particular particle number concentrations, cloud condensation nuclei, $PM_{2.5}$, aerosol optical depth, sulphate and SOA concentrations, cloud reflectivity, liquid water path, precipitation and aerosol ERF. These 9 atmospheric model parameters are considered the most likely causes of uncertainty in low-altitude clouds and aerosol-cloud interactions because they influence boundary layer clouds by altering cloud radiative properties, cloud drop concentrations and microphysical processes, atmospheric humidity, convection processes and boundary layer stability.

A probability density function was defined for each parameter to represent shared expert beliefs about parameter uncertainty. These distributions have no effect on the model simulations (although the ranges define the span of the parameter space), but are used at the stage of generating probability distribution functions (pdfs) of model output based on Monte Carlo sampling from the emulators. We used mainly trapezoidal distributions that avoid overly-centralised Monte-Carlo sampling of the multi-dimensional parameter space (Yoshioka et al., 2018).

Maximin Latin Hypercube sampling was used to create an initial set of 162 simulations that sample model output across the 27-dimensional parameter space. A further set of 54 simulations was used to validate the emulators. In total 217 perturbed parameter simulations of the global model were run for a full year for each anthropogenic emission period (1850, 1978 and 2008 emissions). Each simulation had a spin-up period of seven months from a consistent starting simulation, where the parameters were set to their median values for the first four months and the perturbations then applied in the final three months. Twenty-five simulations did not complete the full annual cycle so were not used in our analysis. Consequently, the ensemble of simulations used for analysis for each period was made up of the remaining 191 simulations, all of which were used to build the final emulators. Radiative forcings were calculated as the difference in top-of-atmosphere (ToA) radiative flux for pairs of simulations with identical parameter settings but different anthropogenic emissions (1850, 1978 and 2008).

Index	Name	Type	Description
1	Rad_Mcica_Sigma	Atm	Fractional standard deviation of sub-grid condensate seen by radiation (controls the overlap of sub-grid clouds)
2	C_R_Correl	Atm	Cloud and rain sub-grid horizontal spatial correlation (determines the accretion rate of cloud drops and aerosols by rain drops)
3	Niter_Bs	Atm	Number of microphysics iteration sub-steps
4	Ent_Fac_Dp	Atm	Entrainment amplitude scale factor (controls the convective mass flux and sensitivity of convection to relative humidity)
5	Amdet_Fac	Atm	Mixing detrainment rate scale factor (controls the rate of humidification of the atmosphere and the shape of the convective heating profile)
6	Dbsdtbs_Turb_0	Atm	Cloud erosion rate per second (The rate at which unresolved sub-grid motions mix clear and cloudy air)
7	Mparwtr	Atm	Maximum value of the function controlling convective parcel maximum condensate

8	Dec_Thres_Cld	Atm	Threshold for the ratio of buoyancy consumption to production before decoupling occurs
9	Fac_Qsat	Atm	Rate of change in convective parcel maximum condensate with altitude
10	Ageing	Aer	Ageing of hydrophobic aerosols (no of monolayers of soluble material)
11	Cloud_pH	Aer	pH of cloud droplets (used to calculate the conversion of SO_2 into sulphate)
12	Carb_BB_Ems	Aer	Carbonaceous biomass burning emissions scale factor
13	Carb_BB_Diam	Aer	Carbonaceous biomass burning emission diameter (nm)
14	Sea_Spray	Aer	Sea spray aerosol scale factor
15	Anth_SO2	Aer	Anthropogenic SO_2 emission scale factor
16	Volc_SO2	Aer	Volcanic SO_2 emission scale factor
17	BVOC_SOA	Aer	Biogenic secondary aerosol formation from volatile organic compounds scale factor
18	DMS	Aer	Dimethyl sulphide surface ocean concentration scale factor
19	Dry_Dep_Acc	Aer	Accumulation mode dry deposition velocity scale factor
20	Dry_Dep_SO2	Aer	SO_2 dry deposition velocity scale factor
21	Kappa_OC	Aer	Kappa-Kohler coefficient of organic carbon
22	Sig_W	Aer	Updraft vertical velocity standard deviation (used to calculate the activation of aerosols into cloud drops)
23	Dust	Aer	Dust emission scale factor
24	Rain_Frac	Aer	Fraction of cloud-covered area in large-scale clouds where aerosol scavenging by rain drops occurs
25	Cloud-Ice_Thresh	Aer	Threshold of cloud ice fraction above which nucleation scavenging is suppressed (restricting further activation of aerosols into cloud drops)
26	BC_RI	Aer	Imaginary part of black carbon refractive index
27	OC_RI	Aer	Imaginary part of organic carbon refractive index

Table 1: The 27 host model and aerosol parameters included in the PPE. Further details are provided in a separate publication (Regayre et al., 2018).

2.4 Model emulation and Monte Carlo sampling

For each model output (such as the regional mean ToA flux, CCN conc., etc. for Europe in July) we construct a statistical emulator model over the 27-dimensional parameter uncertainty using the 137 training simulations and validate it using the 54 validation simulations (as described in Lee et al., 2011). Once validated, a further new emulator is then created using both the training and the validation simulations of the PPE, to obtain a final emulator based on all of the information that our simulations contain. A 'leave-one-out' validation procedure (where each simulation in turn is removed from the merged set, and a new emulator is constructed and used to predict that removed simulation) is applied to additionally verify the quality of our final emulator. We then use this emulator to predict the model output for a large sample (here 4 million) of parameter input combinations that span the 27-dimensional parameter space of the PPE. From this sample we obtain a pdf of the uncertainty in this output variable caused by the defined uncertainty in the model parameters (left hand side of Figure 1). In each case, the output pdf can be sampled according to the elicited parameter probability distributions (Yoshioka et al., 2018), in which case the pdf accounts for prior beliefs about the likelihood of different parameter values. Alternatively uniform sampling can be

applied, in which case the output pdf assumes that all parameters have equal likelihood of lying between their elicited upper and lower limits. Our approach is to use the prior probability distributions, informed by expert knowledge, to sample the parameter combinations of the 4 million model variants over the 27-dimensional parameter uncertainty space.

2.5 Sensitivity Analysis

Sensitivity analysis (Lee et al., 2011; Saltelli et al., 1999) is used to decompose the uncertainty in European regional mean aerosol properties, trends and forcing variables for July into contributions from each individual model parameter. Here, we use the extended-FAST method (Saltelli et al., 1999) in the R package 'sensitivity' (Pujol et al., 2008) to sample from the emulators (as described in section 2.4) and decompose the variance into its individual sources. We then calculate the percentage by which the total variance (for a specific model output) would be reduced if the value of the parameter in question was known precisely. These percentage reductions are used in the analysis of the main causes of model uncertainty in section 3.3.

2.6 Synthetic observations

The 'observations' are taken from the output of the PPE member with each model parameter set to the median value from its corresponding elicited prior distribution. This PPE member was chosen as it lies reasonably centrally within the 27-dimensional parameter uncertainty space. We also tested a marginal set of observations (from a PPE member that had many parameter values located towards the edges of their uncertainty range) but the conclusions of our study did not change, so we focus on the results from the more centralised choice of observations.

We use synthetic observations (Table 2) of European July-mean cloud condensation nuclei (CCN) concentration at 0.2% supersaturation at approximate cloud-base height, surface concentrations of $PM_{2.5}$, mass concentrations of sulphate, OC and BC at the surface, the outgoing shortwave radiative flux at the top of the atmosphere (ToA flux), AOD at a wavelength of 550 nm, and the change in AOD (ΔAOD) and surface solar radiation (ΔSSR) between 1978 and 2008. The period 1978 to 2008 was originally chosen because it is an interesting period for global and regional forcing changes. Although AOD measurements are not available back to 1978, this is not vital to the present study which aims to assess potential constraint over a period with substantial aerosol changes.

The observation uncertainties are based on our judgement about the combined effect of instrument uncertainties and the uncertainty associated with measurement representativeness (colocation of high-frequency point measurements within low-spatial-resolution, monthly-mean model output subject to meteorological variability (Reddington et al., 2017; Schutgens et al., 2016a, 2016b). Where available, we have used sets of real observations to inform these judgements and estimates. For example, we selected our uncertainty range on the ToA Flux such that it is in line with information from the Clouds and the Earth's

Radiant Energy System (CERES) and IPCC uncertainty estimates (Hartmann et al., 2013). In the constraint process we also account for the emulator error (i.e., the estimated uncertainty in each of the 4 million points associated with using the emulator instead of the model itself).

There are other constraints that could be applied to the model such as the aerosol spatial distribution (Myhre et al., 2009), aerosol vertical profile, absorption AOD and single-scatter albedo. It would also be possible to screen the model variants according to skill in capturing high temporal resolution variability (Myhre et al., 2009) or skill in different regions dominated by different aerosols (Shindell et al., 2013). Here, in this idealized constraint exercise, we restrict the analysis to monthly mean aerosol properties over Europe.

Observable Quantity	**Value**	**Uncertainty Range**
Top of atmosphere upward SW flux (W m^{-2})	129	122 – 135
Change in surface downward solar radiation from 1978 to 2008, ΔSSR	3.8	3.2 – 4.4
Cloud condensation nucleus (CCN) conc. at 0.2% supersaturation (cm^{-3})	536	483 – 590
Aerosol optical depth (AOD)	0.17	0.14 – 0.19
Change in AOD from 1978 to 2008, ΔAOD	-0.05	-0.06 – -0.04
$PM_{2.5}$ mass conc. (μg m^{-3})	8.0	7.2 – 8.8
Particle sulphate conc. (μg m^{-3})	1.7	1.2 – 2.2
Particle OC conc. (μg m^{-3})	4.4	3.9 – 4.8
Particle BC conc. (μg m^{-3})	0.23	0.21 – 0.26

Table 2: Observed quantities and corresponding uncertainty ranges used for the constraints applied over Europe. Values are a European July mean, synthetically generated from the model output of a selected PPE member.

2.7 Identification of plausible model variants

Observationally plausible model variants are defined to be those that simulate aerosol and radiation properties within the uncertainty ranges of the observations, defined in Table 2. As we use statistical emulators to generate the simulated output values for each model variant, rather than using the climate model directly, an emulator prediction error φ (valued at one standard deviation on the emulator prediction from the Gaussian process uncertainty) is also taken into account. Hence for a given observed variable, a model variant is rejected as implausible if the range defined by its emulator prediction +/- φ lies

outside the corresponding observation's uncertainty range in Table 2. Furthermore, for a joint observational constraint we retain only the model variants that are classed as plausible for all individual observation types that make up the joint constraint. Such a criterion is possible with synthetic observations because we know that the idealized truth is within the model uncertainty space, but the methodology would be more complex if we were using real observations. It is likely that some of the real observations will deviate from the model significantly because of model structural errors and issues related to the representativeness of the observations. Therefore, the use of real observations would necessitate the definition of a measure of plausibility that accounts for known structural and representativeness errors (McNeall et al., 2016; Williamson et al., 2013).

2.8 Aerosol effective radiative forcing (ERF)

We test the effect of observational constraint on the pre-industrial (PI, here 1850) to present-day (PD, 2008) July-mean European aerosol ERF and its components ERF_{ACI} (Aerosol-Cloud Interaction) and ERF_{ARI} (Aerosol-Radiation Interaction) as well as on the clear-sky component of the ERF_{ARI} (ERF_{ARIclr}). The ERFs (except the ERF_{ARIclr} term) account for above-cloud aerosol scattering and absorption (Ghan, 2013) and are calculated using a fixed sea-surface temperature from 2008.

3 Results

3.1 Relationships among the observed quantities and forcing variables

Figure 2 shows pairwise scatter plots of the PPE member output (Europe July-mean), which provides an overview of the spread of the model outputs as well as the relationships between the variables. We further quantify any linear relationships between the variables using the Pearson correlation coefficient (r) in Table 3.

The aerosol variables show clear inter-relationships. In particular, AOD and $PM_{2.5}$ concentration show the strongest relationship (Pearson correlation, r = 0.88), which is expected given that satellite AOD measurements are frequently used as a proxy for ground-level $PM_{2.5}$ (Chu et al., 2016). This suggests that AOD and $PM_{2.5}$ observations will constrain the model uncertainty to a similar extent and therefore only one of these observable quantities is required. AOD and $PM_{2.5}$ are also clearly correlated with sulphate, OC and BC, which are major components of $PM_{2.5}$ in polluted regions. CCN has a relatively weak positive relationship to both AOD (r = 0.46) and $PM_{2.5}$ (r = 0.21). A positive correlation is expected because, in general, greater aerosol loading will produce greater CCN concentrations, but the correlations are weak because the model aerosol size distribution (which determines CCN) can be configured in many different ways to produce the same AOD. The weak AOD-CCN relation has implications for model constraint: for example, AOD values in the range 0.15-0.2 encompass CCN concentrations of around 400 to around 1000 cm^{-3}. These results are similar to those of (Stier, 2015) who showed similarly weak CCN-AOD correlations.

There are clear relationships between industrial-period forcing variables and some of the observable aerosol properties. For ERF_{ARI} and ERF_{ARIclr} the strongest relationships are with the sulphate concentration (r = -0.77 for ERF_{ARIclr} and -0.67 for ERF_{ARI}) and multi-decadal ΔAOD (r = 0.76 for ERF_{ARIclr} and 0.55 for ERF_{ARI}). As expected, a present-day higher sulphate concentration corresponds to a stronger (more negative) ERF_{ARI}. ΔAOD is negative over Europe due to the reductions in anthropogenic aerosol emissions. Parameter settings that produce a strong multi-decadal ΔAOD also tend to produce a strong pre-industrial to present-day ERF_{ARIclr} and therefore a stronger ERF_{ARI}. Based on these relationships, uncertainty in ERF_{ARIclr} would be easier to constrain than uncertainty in ERF_{ARI} and the most useful aerosol observation for this purpose would be European-mean atmospheric sulphate concentration.

For the ERF_{ACI} there is a relationship with the reflected shortwave ToA flux (r = -0.65), with a larger flux corresponding to a stronger (more negative) forcing. This relationship means that the parameter settings that produce more reflective aerosols and clouds in the present-day atmosphere also enhance ERF_{ACI} forcing. There is also a relationship between the aerosol ERF (preindustrial to present-day) and the 1978-2008 change in surface shortwave radiation (ΔSSR; r = -0.32). However, there is a lot of scatter in the relationship because the model parameters that cause uncertainty in decadal radiative changes are similar but not identical to those that cause uncertainty in forcing over the full industrial period (Regayre et al., 2018, 2014). The relationship between ΔSSR and aerosol ERF among eight models was used by Cherian et al. (2014) as an emergent constraint on aerosol ERF over Europe. In section 4 we explore the implications of our uncertainty analysis for such emergent constraint studies.

In summary, the identified relationships in Figure 2 suggest that for Europe, constraints on sulphate concentration and ΔAOD could lead to some constraint on uncertainty in ERF_{ARI} and ERF_{ARIclr}. Constraint on ToA flux could lead to some constraint on uncertainty in aerosol ERF and ERF_{ACI} over Europe and observed multi-decadal changes in SSR could provide additional constraint. Observational constraints of ERFs are explored in section 3.5.

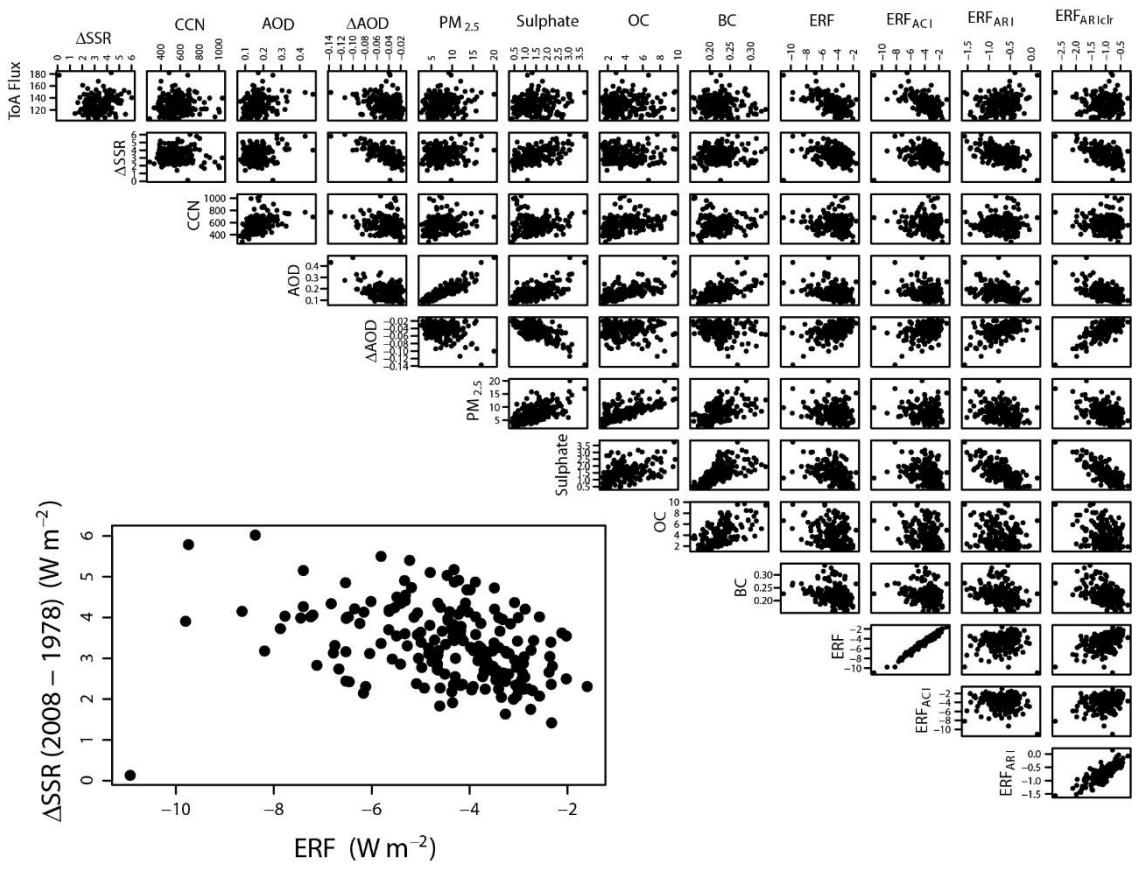

Figure 2. Pairwise scatter plots of the PPE member regional mean model output for Europe in July, for the aerosol properties used as constraints: ToA flux (W m^{-2}), change in SSR (ΔSSR, W m^{-2}) between 1978 and 2008, CCN conc. (cm-3), AOD, surface mass concentrations of $PM_{2.5}$, Sulphate, OC, and BC (μg m^{-3}), the changes in AOD (ΔAOD, W m^{-2}) between 1978 and 2008, and the 1850-2008 forcing variables: aerosol ERF, ERF_{ACI}, ERF_{ARI} and ERF_{ARIclr} (W m^{-2}).

ToA Flux	0.20	0.00	0.23	-0.30	0.16	0.10	0.04	-0.03	-0.59	-0.65	0.25	-0.05
	ΔSSR	-0.04	0.22	-0.72	0.12	0.49	-0.04	0.11	-0.32	-0.22	-0.52	-0.61
		CCN	0.46	-0.20	0.21	0.03	0.37	0.09	-0.14	-0.14	-0.01	-0.09
			AOD	-0.50	0.88	0.59	0.66	0.59	-0.33	-0.27	-0.33	-0.55
				ΔAOD	-0.32	-0.71	-0.17	-0.20	0.48	0.38	0.55	0.76
					$PM_{2.5}$	0.54	0.69	0.54	-0.21	-0.15	-0.33	-0.44
						Sulphate	0.47	0.64	-0.41	-0.28	-0.67	-0.77
							OC	0.62	-0.34	-0.31	-0.20	-0.30
								BC	-0.27	-0.22	-0.28	-0.46
									ERF	0.98	0.16	0.48
										ERF_{ACI}	-0.04	0.33
											ERF_{ARI}	0.83
												ERF_{ARIclr}

Table 3: Pearson linear correlations (r) between the PPE member regional mean model outputs for Europe in July, for the aerosol properties used as constraints, corresponding to the pairwise scatter plots in Figure 2.

3.2 Uncertainty in aerosols and radiative forcings

Figure 3 shows probability density functions of the observable aerosol quantities and the ERFs from the Monte Carlo sample of 4 million model variants (section 2.4). These pdfs sample the complete multi-dimensional parameter space of the model, weighted according the prior probability distributions on the input parameters (Yoshioka et al., 2018). The ranges are similar to those in Figure 2, but the PPE members themselves do not sample the parameter space densely enough to enable a statistically robust pdf to be generated.

The most uncertain observable aerosol properties, with the largest relative standard deviation (ratio of standard deviation to mean value) in our sample are the sulphate and OC concentrations and the multi-decadal ΔAOD (Table 4). This suggests that constraining these properties will substantially reduce the sample size and constrain the parameter space. Some of the pdfs have long tails (e.g. OC concentration and ΔAOD) which suggests that a subset of parameters may be combining in a specific

manner to obtain these extreme values. The tails of the forcing pdfs contain the values most likely to be considered implausible against observations (Regayre et al., 2018).

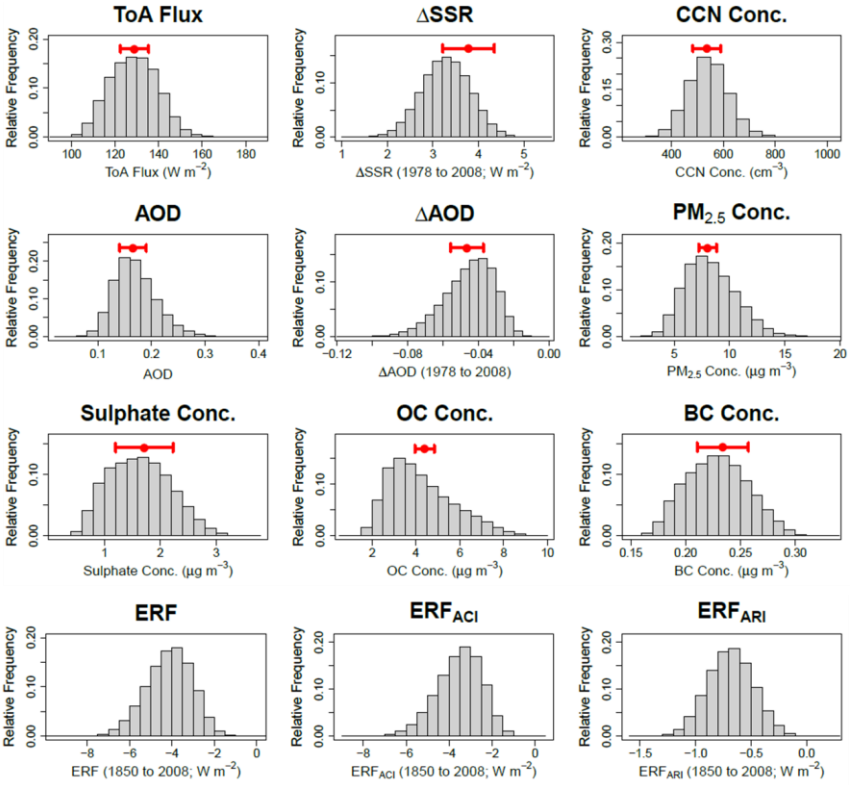

Figure 3. Calculated uncertainty in the aerosol quantities and aerosol ERF terms from the 4 million member sample. Results are for July-mean over Europe. The red bar shows the assumed range of each synthetic observation used to constrain the uncertain parameter space and the aerosol forcing uncertainty from Table 2.

	mean	sd	|sd/mean|
Top of atmosphere upward SW flux (W m^{-2})	128.66	10.83	0.08
Change in surface solar radiation from 1978 to 2008 (W m^{-2})	3.29	0.52	0.16
Cloud condensation nucleus (CCN) conc. at 0.2% supersaturation (cm^{-3})	542.95	78.1	0.14
Aerosol optical depth (AOD)	0.17	0.04	0.23
Change in AOD from 1978 to 2008, ΔAOD	-0.05	0.01	0.31
$PM_{2.5}$ mass conc. (μg m^{-3})	8.27	2.28	0.28
Particle sulphate conc. (μg m^{-3})	1.63	0.54	0.33
OC particle conc. (μg m^{-3})	4.25	1.48	0.35
BC particle conc. (μg m^{-3})	0.23	0.03	0.12
1850 to 2008 ERF (W m^{-2})	-4.14	1.07	0.26
1850 to 2008 ERF_{ACI} (W m^{-2})	-3.52	1.04	0.3
1850 to 2008 ERF_{ARI} (W m^{-2})	-0.68	0.2	0.3
1850 to 2008 ERF_{ARIclr} (W m^{-2})	-1.02	0.28	0.27

Table 4: Mean, standard deviation (sd) and absolute relative standard deviation (|sd/mean|) of the calculated uncertainty in the aerosol quantities and aerosol ERF terms from the 4 million member sample. Results are for July-mean over Europe.

3.3 Sensitivity Analysis

We can decompose the overall variance in a model output into percentage contributions from the individual input parameters (Lee et al., 2011; Saltelli et al., 1999). The results of this analysis are shown in Figure 4.

For many of the output variables there is little correspondence with the forcing variables in terms of the main parameters that cause uncertainty. In particular, only about 10% of the CCN concentration uncertainty comes from the main causes of uncertainty in any of the corresponding forcing variables, which is consistent with the weak correlations in Figure 2 and the conclusions of Lee et al. (2016).

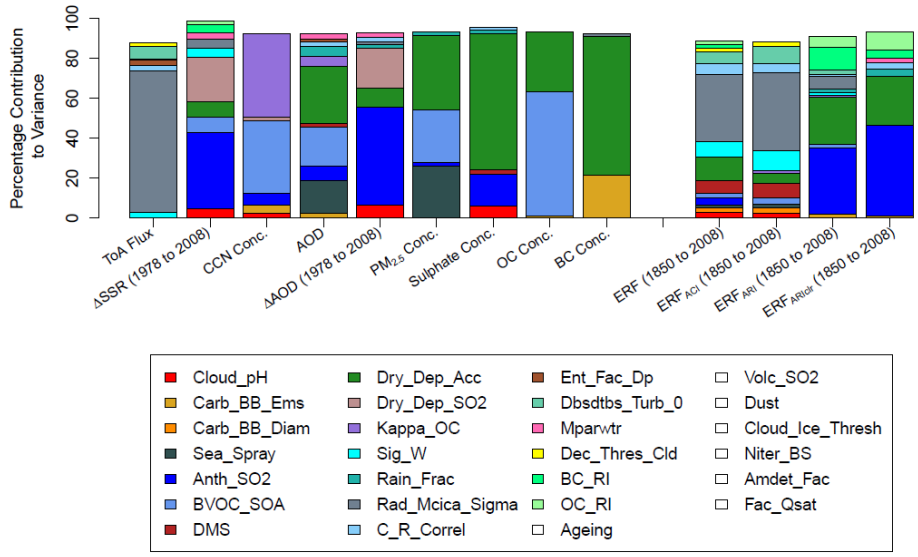

Figure 4. Variance-based sensitivity analysis results, showing the percentage parameter contributions to model output uncertainty in the observable aerosol quantities and the forcing variables for Europe. Only those parameters which cause at least 1% of the variance are shown in colour.

There is reasonable correspondence between the sources of uncertainty in sulphate concentration, ΔAOD, ERF_{ARI} and ERF_{ARIclr}, which is again consistent with Figure 2. Around 60-70% of the output variance in these variables is accounted for by anthropogenic SO_2 emissions (Anth_SO2) and the accumulation mode dry deposition velocity (Dry_Dep_Acc). This degree of similarity in the parametric uncertainty sources implies that an individual observational constraint on ΔAOD should lead to some constraint of the ERF_{ARI} and ERF_{ARIclr} forcing uncertainty. Dry_Dep_Acc is also a significant cause of uncertainty in AOD, $PM_{2.5}$ and concentrations of sulphate, OC and BC (between ~25% and ~70% for each). Hence, it is possible that constraint of these observable aerosol quantities may lead to some constraint on ERF_{ARI} and ERF_{ARIclr} uncertainty. We also see that the uncertainty in the ToA flux is dominated by the cloud radiation parameter Rad_Mcica_Sigma parameter (which affects the spatial homogeneity of the clouds), which also accounts for about 35% of the variance in ERF and ERF_{ACI}. This parameter also causes most of the uncertainty in global mean ERF and ToA flux (Regayre et al., 2018). However, over Europe there are multiple other parameters causing a small amount of the aerosol ERF uncertainty which suggests an effective constraint will require using multiple complementary observations.

3.4 Constraint of aerosol properties

We first explore how constraining an individual aerosol property helps to constrain the range of other observable properties and multi-decadal trends. AOD is the aerosol property most frequently observed and used to evaluate and constrain models (e.g., Shindell et al. (2013)) and is used as the control variable in data assimilation used to evaluate the aerosol forcing (Bellouin

et al., 2013). Figure 5 shows the reduction in uncertainty of the modelled atmospheric properties and trends by constraining the model to match observed AOD. Credible intervals (95%) corresponding to the individual constraints are provided in Table 5.

Constraint of European monthly-mean AOD to lie in the range 0.14-0.19 (23% of the full ensemble range) leads to a fairly strong constraint of $PM_{2.5}$ uncertainty: the standard deviation of $PM_{2.5}$ ($\sigma PM_{2.5}$) is reduced by 34% when the range of AOD is reduced by about 77%. The standard deviation of the PM components and the multi-decadal trend ΔAOD are also reduced, but by a smaller amount: around 20% for OC and only around 10% for BC, sulphate and ΔAOD. Constraint of individual chemical components is weaker because there are many combinations of sulphate, BC and OC that can account for high or low AOD. Uncertainty in the other observable quantities (CCN and ToA flux and the multi-decadal trend ΔSSR) are essentially unaffected by the constraint of AOD. The reason for the weak constraint is that there are many model variants within the observed range of AOD (or $PM_{2.5}$) that produce very different CCN, ToA flux and ΔSSR.

These results provide some indication of the possible remaining uncertainty in a model that has been tuned to agree with AOD observations. A tuned model that agrees with AOD observations within the observational uncertainty is just one of many potential variants of the model that have equally good agreement with the observations. For example, our model suggests that the remaining uncertainty (absolute range) in European-mean CCN could be 755 cm^{-3} in a model constrained by AOD observations, which is only slightly less than the unconstrained range of 782 cm^{-3}. Most surprisingly, constraint of AOD leaves open a wide range of potential values of the change in AOD over decadal periods. The range of ΔAOD from 1978-2008 after constraint of 2008 AOD is 0.105 (range -0.109 to -0.004), which is only slightly lower than the unconstrained range of 0.115. Screening model variants based on their ability to reproduce a single aerosol-related observation is not a sufficient constraint on aerosol-related model uncertainty. Therefore tuning a model to AOD observations is completely inadequate for producing a robust aerosol model.

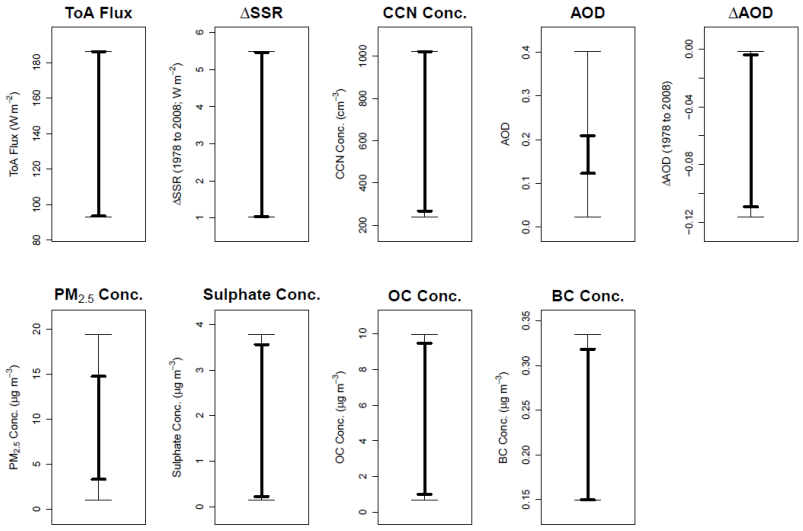

Figure 5. Effect of observational constraint of AOD on other aerosol properties in the model over Europe. The bars show the absolute range of the pdf before (thin line) and after (thick line) constraint. Results are for mean properties over Europe in July.

5

	95th CI unconstrained	**95th CI constrained by AOD**
Top of atmosphere upward SW flux (W m^{-2})	(108.9 , 149.5)	(109.0 , 149.0)
Change in surface solar radiation from 1978-2008 (W m^{-2})	(2.27 , 4.30)	(2.29 , 4.30)
Cloud condensation nucleus (CCN) conc. at 0.2% supersaturation (cm^{-3})	(396 , 704)	(408 , 698)
Aerosol optical depth (AOD)	(0.105 , 0.257)	(0.130 , 0.201)
Change in AOD from 2008 – 1978, ΔAOD	(-0.076 , -0.022)	(-0.072 , -0.024)
$PM_{2.5}$ mass conc. (μg m^{-3})	(4.41 , 13.15)	(5.31 , 10.98)
Particle sulphate conc. (μg m^{-3})	(0.71 , 2.72)	(0.79 , 2.58)
OC particle conc. (μg m^{-3})	(2.10 , 7.61)	(2.24 , 6.66)
BC particle conc. (μg m^{-3})	(0.179 , 0.284)	(0.183 , 0.277)

Table 5. Effect on the uncertainty in aerosol properties over Europe when the model is constrained by $PM_{2.5}$ measurements (assumed measurement uncertainty range 7.2-8.8 μg m^{-3}) or AOD (assumed measurement uncertainty range 0.14-0.19). The aerosol uncertainties are given as the 2.5th and 97.5th empirical percentiles of the pdf to form a 95% credible interval.

3.5 Constraint of 1850 – 2008 aerosol ERF uncertainty

3.5.1 Effects of individual aerosol and radiation observational constraints

Figure 6 (top row) shows the potential constraint achievable on uncertainty in the 1850–2008 aerosol ERF, ERF_{ACI}, ERF_{ARI} and ERF_{ARIclr} when we constrain July-mean AOD over Europe. Each box and whisker plot shows the uncertainty distribution from the original sample of 4 million model variants (grey, left) and the sample of constrained models (pink, right). Table 6 shows means and standard deviations for the original and constrained distributions from AOD and all other individual observational constraints.

Observational constraint of simulated AOD has essentially no effect on the range of aerosol ERF and the ERF_{ACI} component of forcing over Europe. There is some reduction in uncertainty in the ERF_{ARIclr} component of forcing (standard deviation reduced by around 12%) but not in ERF_{ARI}, despite both sharing common causes of uncertainty with AOD (Section 3.3).

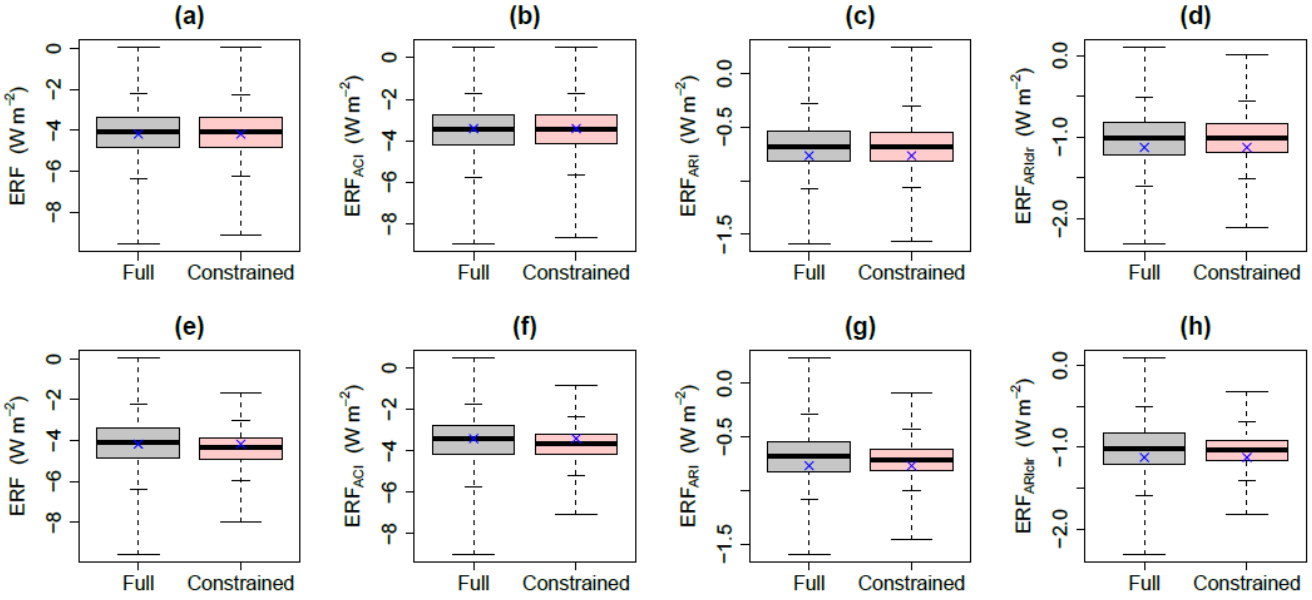

Figure 6. The effect of the AOD constraint (top row) and all observational constraints together (bottom row) on the uncertainty in the 1850-2008 forcing variables over Europe. Columns left to right show constraint of aerosol ERF, ERF_{ACI}, ERF_{ARI} and ERF_{ARIclr} (W m^{-2}) respectively. The boxes show the inter-quartile range (with the median value shown by the black line that cuts it) and the whiskers show the full range of the distribution. The short horizontal bars on the whiskers correspond to 95% credible interval bounds. The grey boxes show the distribution for the variable predicted over the original sample (4 million model variants that span the underlying parameter uncertainty) and the pink boxes show the corresponding distribution of the remaining samples after the constraint has been applied. The predicted forcing using the input combination of the model run used as the idealized observation is shown by the blue cross.

Applied constraint	ERF (W m^{-2})	ERF_{ACI} (W m^{-2})	ERF_{ARI} (W m^{-2})	ERF_{ARIclr} (W m^{-2})
No constraint	-4.144 (1.075)	-3.523 (1.044)	-0.683 (0.204)	-1.024 (0.281)
All constraints	-4.391 (0.759)	-3.707 (0.736)	-0.710 (0.148)	-1.044 (0.184)
ToA Flux	-4.137 (0.863)	-3.496 (0.790)	-0.683 (0.201)	-1.024 (0.280)
ΔSSR (1978-2008)	-4.247 (1.045)	-3.593 (1.034)	-0.710 (0.193)	-1.067 (0.260)
CCN Conc.	-4.181 (1.065)	-3.547 (1.036)	-0.695 (0.204)	-1.039 (0.277)
AOD	-4.123 (1.034)	-3.500 (1.017)	-0.684 (0.194)	-1.014 (0.246)
ΔAOD (1978-2008)	-4.175 (1.003)	-3.541 (1.011)	-0.693 (0.176)	-1.034 (0.214)
$PM_{2.5}$ Conc.	-4.173 (1.057)	-3.541 (1.039)	-0.688 (0.197)	-1.021 (0.257)
Sulphate Conc.	-4.231 (1.037)	-3.589 (1.040)	-0.695 (0.167)	-1.035 (0.222)
OC Conc.	-4.245 (1.047)	-3.622 (1.027)	-0.682 (0.207)	-1.018 (0.278)
BC Conc.	-4.263 (1.058)	-3.607 (1.046)	-0.707 (0.195)	-1.052 (0.259)

Table 6: Mean and standard deviation (in brackets) of the forcing distributions over Europe for the original unconstrained sample, the multiple-constraint sample and the individually constrained samples where each observational constraint is applied independently.

Figure 7 summarises the effect of the other individual constraints. For ERF_{ACI} (and therefore aerosol ERF, which is dominated by ACI) the only observation that has any meaningful effect on the range is the ToA flux. When the flux is constrained to be within the range 122-135 W m^{-2} (from the prior range of 90-175 W m^{-2}) the standard deviation of ERF_{ACI} over Europe falls by 24% (Table 6). The ΔSSR observation reduces the aerosol ERF and ERF_{ACI} standard deviations by less than 3%. The only other constraints on uncertainty in aerosol ERF and ERF_{ACI} come from constraining AOD or ΔAOD, which reduce the forcing uncertainties by around 3% each. The effect of applying all observations in combination is discussed in Section 3.5.2.

ERF_{ARI} and ERF_{ARIclr} are constrained by several individually applied observations. ΔAOD and sulphate concentrations provide the strongest constraints. ΔAOD reduces the standard deviation of ERF_{ARI} and ERF_{ARIclr} by 14% and 24% respectively. Constraining sulphate concentrations reduces the uncertainty in ERF_{ARI} by 18% and in ERF_{ARIclr} by 21%. The strong constraint of ERF_{ARI} and ERF_{ARIclr} uncertainty is consistent with Figure 4, where we saw that around 60-70% of the uncertainty in each of ΔAOD, ERF_{ARI} and ERF_{ARIclr} could be attributed to the same two parameters. Again, the relatively weak constraint is caused by interacting combinations of parameter effects ((Lee et al., 2016; Regayre et al., 2018); Section 3.7), so there is potential for

significant error compensation (or equifinality, Beven and Freer, 2001). In all other cases the individual observational constraints reduce the uncertainty in ERF_{ARI} and ERF_{ARIclr} by less than 10%.

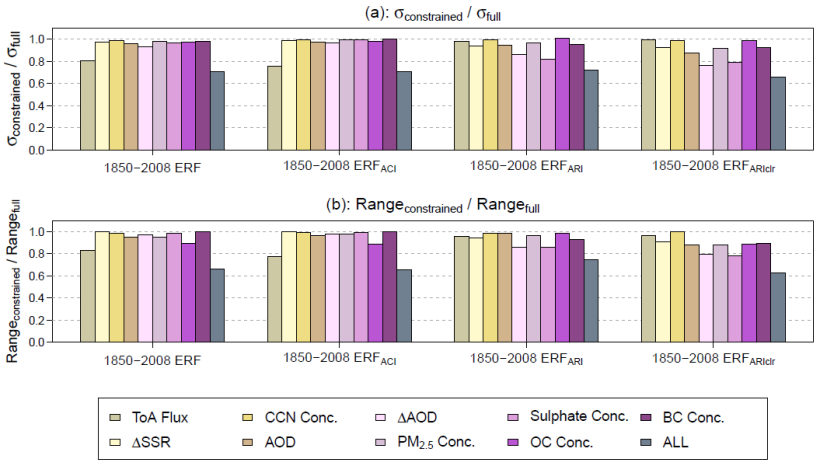

Figure 7. The relative constraint achieved for aerosol ERF, ERF_{ACI}, ERF_{ARI} and ERF_{ARIclr} over Europe given the individual synthetic constraints applied (colours) and the simultaneous constraint (ALL). The relative constraint is evaluated as the ratio of the standard deviation of the forcing in the constrained sample ($\sigma_{constrained}$) to the standard deviation of the forcing in the original, unconstrained sample (σ_{full}).

3.5.2 Effect of all observational constraints

Figure 6 (bottom row) and the right-most bars in Figure 7 show the reduction in Europe-mean 1850 – 2008 aerosol ERF, ERF_{ACI}, ERF_{ARI} and ERF_{ARIclr} uncertainty when we simultaneously apply all nine observational constraints. The standard deviations are reduced by 29.4% for the aerosol ERF, 29.5% for ERF_{ACI}, 27.8% for ERF_{ARI} and 34.3% for ERF_{ARIclr} (Table 6) and Figure 6 shows a reduction in the interquartile range (box width) and 95% credible interval in each case.

Our results show that multiple observational constraints are very effective at reducing the plausible parameter space (ruling out 96.4% of model variants). However, these reductions in parameter space have only a modest impact on aerosol ERF uncertainty. This occurs because the 27 parameter values in the constrained space can be combined to produce a wide range of ERFs (Lee et al. 2016). These results highlight the value of exploring the wider underlying modelling uncertainties (achieved here using a well-designed PPE to inform statistical emulation). The more comprehensive exploration of the parameter space using several million model variants from the emulators enabled us to explore the wider uncertainties that would not have been captured even by the 191 PPE members. Furthermore, a 96.4% reduction in parameter space would have reduced the number of PPE members to one or two, which would not have revealed that a large fraction of the ERF uncertainty (70.6%) remained unconstrained. Likewise, a single model variant arrived at through tuning cannot represent model behaviour over the remaining

plausible parameter space. Similar concerns about non-robust samples apply also to the small number of members in multi-model ensembles.

3.5.3 Effect of combinations of observational constraints

An important question in model constraint is how quickly the model uncertainty falls as additional observational constraints are applied. Figure 8 shows the average reduction in forcing uncertainty versus the number of observational constraints applied. With nine possible observational constraints there are nine possible single constraints, 36 possible combinations of 2 constraints, 252 combinations of 3 constraints, etc. We therefore show a mean over all possible combinations of each number of constraints.

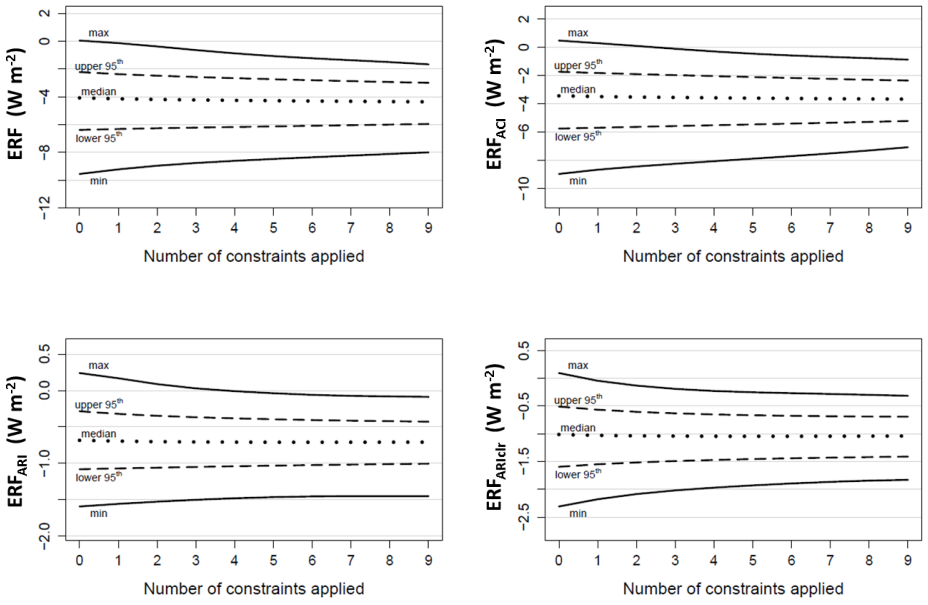

Figure 8. The dependence of aerosol ERF uncertainty on the number of observational constraints applied. The lines show the mean effect of different numbers of constraints.

Averaged across the many combinations of constraints, uncertainty in aerosol ERF and its components initially falls approximately linearly with the number of constraints applied, but then flattens out. This dependence implies that some observations are constraining the same sources of uncertainty as other observations (as shown in section 3.3). So while a large number of observations are needed to constrain forcing, it is also important to identify observations that provide unique constraints on parameter uncertainties. The effectiveness of each observational constraint depends on which other constraints are applied with it. For example, two positively correlated observations like $PM_{2.5}$ and AOD (Figure 2) will reduce the

allowable parameter space in broadly the same dimensions because the same parameters cause their uncertainty (Figure 4). Therefore the constraint on forcing uncertainty achieved by AOD and $PM_{2.5}$ is not additive.

3.6 Constraint of 1978 – 2008 forcing uncertainty

Previous research has shown that the causes of uncertainty in recent decadal forcing are quite different to the causes of uncertainty in pre-industrial to present-day forcing (Regayre et al., 2018, 2014). Much of the uncertainty in PI to PD aerosol–cloud interaction forcing is caused by natural aerosols (Carslaw et al., 2013; Carslaw et al., 2017), which are much less important over recent decades. We therefore expect recent aerosol and radiation observations to provide a greater constraint on recent decadal forcings than on forcing referenced to PI conditions. Our results show that this hypothesis is correct: simultaneous application of the nine observational constraints reduces the standard deviation of the 1978-2008 aerosol forcing distributions by 33.7% for ERF, 32.3% for ERF_{ACI}, 35.0% for ERF_{ARI} and by 43.9% for ERF_{ARIclr}, which are all greater reductions than for the 1850 to 2008 forcing (Figure 7, Table 6). The main contributor to the reduction in uncertainty in the aerosol ERF from 1978 to 2008 is the change in AOD, followed by present-day AOD. These results suggest that forcing uncertainty in recent decades may be more readily constrained by observations than multi-century forcing.

3.7 Constraint of plausible parameter ranges

The overall objective of our approach is to identify all the observationally plausible variants of the model so that they can be used to calculate an observationally constrained spread of aerosol ERFs. Each variant is associated with a particular part of parameter space. We can then use the emulators to compute the constrained magnitude and range of any other aerosol property (or the changes between 1850, 1978 and 2008). Alternatively a sample of these variants (parameter settings) could be used in the model itself to simulate aerosol effects for any situation (for example, with very different meteorological conditions, or anthropogenic aerosol emissions).

Figure 9 shows a one-dimensional projection of the remaining parameter space after constraining to the nine observations. There are some substantial reductions in the plausible marginal range of several individual parameters. It needs to be borne in mind that, with 27 parameter dimensions, the parameter relationships which have been constrained by multiple observations cannot be seen in the one-dimensional projection. That is, the remaining plausible individual parameter values can be combined in many ways with the remaining space of the other parameters and still reproduce all of the observations (Lee et al., 2016; Regayre et al., 2018). Figure 9 identifies parts of the marginal parameter space that are effectively ruled out in white. For example, a very low setting of the BVOC_SOA parameter cannot produce observationally plausible results when combined with any of the possible combinations of the other 26 parameters.

The constraint of the parameter ranges will be different when using real observations, but it is interesting to see how nine observations can marginally constrain twenty-seven parameters when there is a high degree of potential compensating effects. The strongest marginal constraint is on: the sea spray aerosol emissions (Sea_Spray; the highest 25% and lowest 15% of values are implausible), biogenic secondary aerosol formation (BVOC_SOA; the lowest 40% and top 20% of the range are implausible), the hygroscopicity of organic carbon (Kappa_OC; the top 40% of values are implausible), and the imaginary part of organic carbon refractive index (OC_RI; top 30% is implausible). Furthermore, the lowest 10-20% of the range of several aerosol emission parameters are also deemed implausible (biomass burning (Carb_BB_Ems), degassing volcanic (Volc_SO2), DMS, anthropogenic sulphur dioxide (Anth_SO2)).

The atmospheric (host model) marginal parameter ranges are much less constrained because the observable variables that we used are not strongly dependent on them, except for ToA flux observations which are known to be affected by the Dec_Thresh_Cld and Rad_Mcica_Sigma parameters (Regayre et al., 2018). Values of the threshold for cloudy boundary layer decoupling parameter (Dec_Thresh_Cld) are concentrated towards the lower end of the range (the upper 40% are implausible). We also show that the top 20% of values are implausible for the parameter controlling the amount of overlap between sub-grid clouds as seen by the model's radiation code (Rad_Mcica_Sigma). However, the lowest 40% of this parameter range can be entirely ruled out by constraining the ToA flux in the North Pacific (Regayre et al., 2018). These results highlight the important benefits which will come from constraining the model uncertainty using multiple observations in multiple environments.

This analysis highlights the complexity of the multi-dimensional parameter uncertainty space that remains after observational constraint: there are clearly a large number of ways of tuning a model to be observationally plausible.

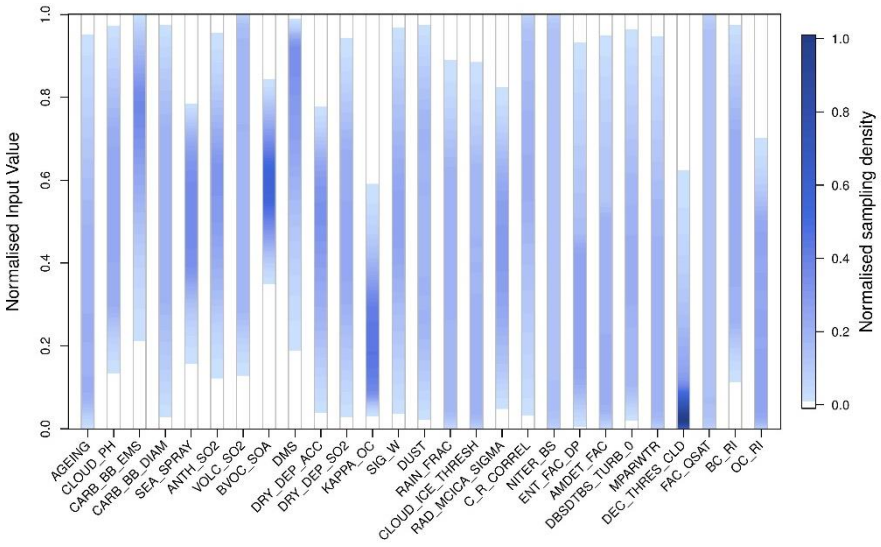

Figure 9. One-dimensional projection of the remaining parameter space after simultaneous constraint of all atmospheric quantities and decadal trends. The colour-scale shows the marginal normalised sampling density (normalised across parameters) of each input parameter over its range. Parts of the marginal parameter space that are effectively ruled out are shown in white (normalised sampling density <0.02).

4 Implications for model screening and emergent constraints

In multi-model ensembles it is usually the case that each modelling group submits a single well-tuned version (variant) of a model. The uncertainty in the ensemble is determined by the structural differences between the models, but the uncertainty in the individual models (caused by multiple uncertain parameter settings) is not quantified. Here we use the uncertainty in HadGEM3-UKCA to estimate the effect it might have on the results of multi-model emergent constraint studies. Clearly the uncertainties in each model will differ, so we use our model uncertainty only as a rough estimate of the potential effect.

In the ACCMIP study (Shindell et al., 2013) model skill at simulating AOD was used to screen nine models. We have described in Sections 3.4 and 3.5 why constraint of AOD can only be considered the first step in model screening; AOD does not effectively reduce model uncertainty when used in isolation. The standard deviation of the modelled global annual mean ERF_{ARI} in the ACCMIP study was about 50% of the multi-model mean. In our results, after we have screened out model variants that are inconsistent with synthetic AOD observations (i.e., effectively tuning to AOD), the standard deviation of the HadGEM-UKCA ERF_{ARI} over Europe is about 30% of our mean. Therefore the standard deviation in HadGEM3-UKCA caused by uncertain input parameters is a significant fraction of the multi-model standard deviation, and would affect the constrained range of ERF_{ARI}. Shindell et al. (2013) acknowledged that uncertainties in the emissions could alter the relative agreement of the models with observations and thereby affect the spread of plausible model predictions. However, uncertainty in emissions is just one of many possible sources of uncertainty that could affect the conclusions (Figure 4).

In emergent constraint studies a linear relationship between aerosol forcing and an observable variable simulated by multiple models is used to define an observationally constrained value of the variable of interest. In the Cherian et al. (2014) study Europe-mean aerosol ERF was estimated by regressing modelled ERF against the 1990-2005 modelled trend in SSR over Europe from seven aerosol-climate models. An observed SSR trend of -4.0 ± 0.6 W m^{-2} $decade^{-1}$ enabled the Europe-mean aerosol ERF to be constrained to -3.56 ± 1.41 W m^{-2} (corresponding to the range of -4.97 to -2.15 W m^{-2}). Their analysis accounted for the uncertainty in SSR caused by meteorological variability but did not account for the influence of parametric uncertainties.

In our HadGEM3-UKCA simulations the Europe-mean aerosol ERF 95% credible interval is -6.0 to -2.7 W m^{-2} after constraining the model using nine observations (i.e., a tight tuning of the model). This range provides some measure of the range of alternative ERFs that could be obtained by the individual models had they been tuned differently (*but equally well*) to observations (although we do not know what actual tuning was undertaken). Our single-model uncertainty range is comparable to the multi-model ensemble range, but was not accounted for by Cherian et al. (2014) in deriving the emergent constraint. The effect of including this previously neglected source of single-model uncertainty is to substantially increase the uncertainty on the emergent constraint (Figure 10). Furthermore, the likely magnitude of forcing derived from emergent constraints is sensitive to the uncertainties accounted for in the process (Samset et al., 2014).

In many emergent constraint studies, the constrained ERF (or other quantity) is essentially based on the very small number of models that lie within the uncertainty range of one observation (Figure 10). With our approach, model variants that are plausible against this one observation type are then examined to determine their plausibility against many other observation types – in this study, nine observations in total. Ultimately, multivariate constraint is essential to reach robust conclusions because of the many compensating sources of model uncertainty.

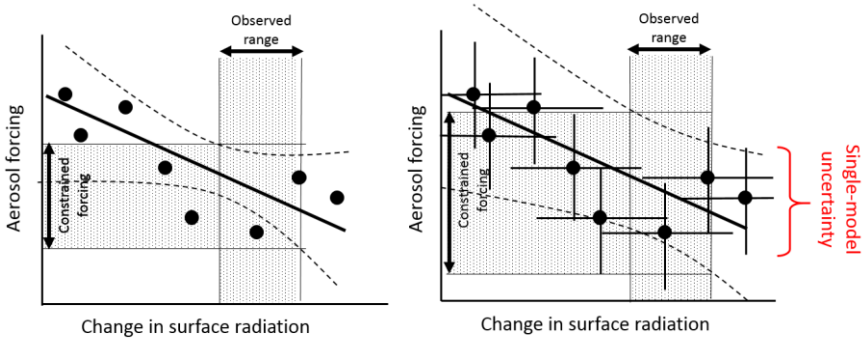

Figure 10. An example of an emergent constraint of the aerosol forcing using results from multiple models. (a) a relatively tight apparent constraint when the uncertainty in the individual models is neglected; (b) a much weaker constraint when the uncertainty in the individual models is accounted for.

5 Conclusions

The use of observations to produce a well-configured model variant is a fundamental aspect of ensuring that models can make trustworthy predictions. For example, in a review of progress on reducing uncertainty in direct radiative forcing Kahn (2012) argues that models can be "constrained by the aggregate of observational data, to calculate the regional and global radiation fields and material fluxes with adequate space–time resolution to produce the best result we can achieve." The primary objective of our study was to determine how much uncertainty could remain in an aerosol-climate model when it is constrained to match combinations of observations that define the base state of the model: top-of-atmosphere upward shortwave flux, aerosol optical depth, $PM_{2.5}$, cloud condensation nuclei, concentrations of sulphate, black carbon and organic material as well as multi-decadal change in surface shortwave radiation and aerosol optical depth. Our results refer to July-mean conditions over Europe.

To estimate the uncertainty that might typically exist in a climate model before and after tuning, we used a perturbed parameter ensemble that sampled uncertainties in 27 parameters related to aerosol emissions, aerosol and cloud processes, and parameters in the host physical climate model that influence clouds, humidity, convection and radiation in the base state of the model. We performed 191 model simulations that spanned the 27-dimensional space of the model uncertainty and then built surrogate model emulators from which we created a Monte Carlo sample of 4 million 'model variants'. Using synthetic observations (taken from the output of one of our simulations) we determine the extent of the potential constraint that these nine aerosol and cloud-related properties can generate. Constraining the model outputs using all nine observations rules out over 96% of the model variants and the associated implausible parts of parameter space. The remaining 153,000 model variants have been used to estimate the observationally constrained aerosol ERF and the uncertainty associated with one tuned version of HadGEM3-UKCA.

Tuning HadGEM3-UKCA to AOD alone has almost no effect on the reliability of the tuned model to simulate CCN (and hence cloud drop number concentrations) (Figure 5). Constraining European-mean AOD in July to lie within a realistic range of 0.14 – 0.19 (23% of the full model uncertainty range) results in a reduction of less than 5% in the uncertainty in CCN generated by the full set of 4 million variants of the model. The CCN range is then 268-1022 cm^{-3} compared to 241 – 1022 cm^{-3}. This provides a measure of the parametric uncertainty when AOD measurements are used to infer CCN, although the range would potentially be larger had we perturbed more parameters (Yoshioka et al., 2018). Tuning a model to AOD alone also has very little effect on the modelled range of the trend in AOD over a multi-decadal period. The complete set of model variants produces a range of 1978 to 2008 changes in AOD over Europe of 0.115, and this range is only reduced to 0.105 (range –0.109 to 0.004) when the model is constrained to match European-mean 2008 AOD within observational uncertainty.

Constraint of AOD alone has a negligible effect on the uncertainty in the aerosol ERF over Europe. Although the aerosol ERF simulated by a model will change as parameters are tuned to achieve agreement with AOD measurements, any resulting ERF will have large uncertainty (i.e., there are many other equally well-tuned model variants that produce different ERFs). This uncertainty cannot easily be estimated without a full uncertainty analysis of the model as we have done here. The weak constraint calls into question the robustness of estimates of aerosol forcing based on AOD reanalyses (Bellouin et al., 2013).

It is often argued that AOD is a poor variable to use for understanding aerosol-cloud interactions. However, our results show that even the most strongly related measurement (CCN) also does not provide a strong individual constraint on ERF_{ACI} (Figure 7). It is doubtful that other derived variables like aerosol index will be any better. The key to model constraint is to find combinations of observations that help to constrain ERF: Individual constraints are unlikely to be effective, although they may appear to be effective if the model uncertainty is not fully sampled.

Observational constraint using nine observations has the potential to reduce the uncertainty in aerosol ERF slightly more over a multi-decadal period than over the full industrial period: the standard deviation falls by 29.4% for the 1850-2008 aerosol ERF, 29.5% for ERF_{ACI}, 27.8% for ERF_{ARI} and 34.3% for ERF_{ARIclr}. The standard deviation of 1978-2008 aerosol ERF could be reduced by around 34%, which is greater than for the 1850-2008 ERF because there is greater correspondence between the causes of uncertainty in near-term aerosol forcing and the 2008 aerosol-cloud-radiation state than there is between the 1850-2008 ERF and the 2008 state (Regayre et al., 2018, 2014). Because near-term future changes in aerosols and clouds are likely to resemble recent changes more than centennial-scale changes, we are optimistic that the uncertainty in near-term aerosol ERFs could be constrained and used to provide policy-relevant information on near-term temperature changes (Hawkins et al., 2017). A shift of emphasis of the research community towards trying to constrain decadal forcing uncertainty, instead of industrial era forcing, is likely to accelerate progress.

The most effective observational constraint on the uncertainty in aerosol ERF and ERF_{ACI} is the top-of-atmosphere upward shortwave flux. When the flux is constrained to be within the range 122-135 W m^{-2} (from the prior range of 90-175 W m^{-2}) the standard deviation of ERF_{ACI} over Europe falls by 24%. Other observational constraints reduce the ERF_{ACI} uncertainty by a few percent at most, including the change in surface SW radiation. Effectively, this result means that routine tuning of radiative fluxes in climate models will have a bearing on the magnitude of the aerosol ERF that the models calculate. The reason for the constraint on forcing uncertainty is that model parameters that control cloud and atmosphere brightness also control how that brightness responds to changes in aerosols over Europe. However, it is only likely to be an effective constraint where the brightness is controlled by tuning the parameters we have identified here. In regions dominated by quite different processes, like mixed-phase clouds, tuning the flux will have a much weaker effect on the aerosol ERF.

The most effective observational constraints on the uncertainty in ERF_{ARI} and ERF_{ARIclr} over Europe are the sulphate concentration and the change in AOD over a multi-decadal period (we used 1978-2008). When applied individually, sulphate concentrations constrain ERF_{ARI} standard deviation in our ensemble by 18% over Europe. The 1978-2008 change in AOD

constrains the ERF_{ARI} standard deviation by 14% when applied individually. Constraint of AOD itself (in 2008) reduces the ERF_{ARI} uncertainty by only 5%, and would not provide a realistic way of screening models. The other constraints were much less effective.

The plausible ranges of some natural aerosol emissions are reduced after constraining to the nine observations, particularly sea spray emissions and biogenic volatile organic aerosol formation. We were also able to constrain some aerosol process parameters such as the CCN hygroscopicity (kappa), the imaginary refractive indices of BC and OC, and parameters controlling boundary layer stability and the radiative properties of overlapping sub-grid clouds which control cloud brightness. Observational constraint generates a set of constrained parameter settings that can be taken forward and used to make model predictions under any other conditions (e.g. for future projections).

The range and combinations of observationally plausible parameter values remain very large even after constraint using nine observations, which explains why the aerosol ERF uncertainty remains large after constraint. This result is not a failure of our approach, but rather an indication of the multiple ways in which uncertain model parameters can combine to predict a wide range of outputs with equal skill when compared to observations. These multiple model variants are neglected when a single model variant is produced through tuning.

Widely used procedures of aerosol-climate model evaluation and observational 'validation' lack statistical robustness because they do not adequately sample the model uncertainty space. We showed that observational constraint against nine observations identified less than 4% of the 4 million sampled points in multi-dimensional parameter space as plausible (i.e., the model value is within the observational uncertainty). A 96% reduction in parameter space would have reduced our original set of 191 ensemble members to one or two, which would not have revealed that a large fraction of the ERF uncertainty (about 71%) remained unconstrained. This creates a fundamental problem for multi-model ensembles (which have far fewer members) and model tuning (which may explore only a few dozen model variants and mostly with single parameter perturbations). From such small samples of models it is not possible to determine how observations help to reduce model uncertainty, so estimates of radiative forcing should not be considered robust.

Our results have implications for studies that seek emergent constraints on a small set of models based on one observational metric. An emergent constraint can be informative, but cannot be expected to reduce the uncertainty in a complex model when used in isolation. The example closest to our study is Cherian et al. (2014) in which the relationship between aerosol ERF and the trend in surface solar radiation (SSR) over Europe for seven climate models was used to estimate the observationally constrained uncertainty in aerosol ERF. Our results for the HadGEM3-UKCA model show that the uncertainty in aerosol ERF and SSR trend in any one tuned version of the model is likely to be of the same order of magnitude as the multi-model range. If the uncertainties in individual models in an ensemble are not accounted for, then we risk being over-confident in the emergent

constraints. Efforts to quantify and observationally constrain individual models are therefore not an alternative to multi-model studies, but individual model uncertainty needs to be quantified and incorporated as an essential component of such efforts to understand and then reduce aerosol ERF uncertainty.

There is considerable scope to extend our approach to incorporate more observation types and more regions. These should include: 1) Aerosol and radiation trends (Allen et al., 2013; Cherian et al., 2014; Leibensperger et al., 2012; Li et al., 2013; Liepert and Tegen, 2002; Shindell et al., 2013; Turnock et al., 2015; Zhang et al., 2017). So far we used changes in AOD and SSR, but changes in ToA SW flux as well as aerosol components like OC (Ridley et al., 2018) and sulphate could provide useful constraints. 2) Observations from pristine regions that might provide a constraint on preindustrial-like aerosol and cloud properties (Carslaw et al., 2017; Hamilton et al., 2014). 3) Information on the vertical profile of aerosols. 4) Observed relationships between changes in aerosol and cloud variables (Ghan et al., 2016; Gryspeerdt et al., 2017; Quaas et al., 2009) such as defined in Eq. 1. Such relationships are a favoured way to constrain forcing. Although it is conceivable that relationships between change-of-state variables can be predicted more reliably than state variables themselves (because of cancellation of correlated model errors), the model uncertainty in these relationships has not been determined in studies that have applied them.

Whichever approach is used to reduce uncertainty in aerosol forcing, it is essential to acknowledge that aerosol-chemistry-climate models are highly complex with dozens of sources of uncertainty that can be combined in many ways. Such a system cannot be constrained by one or two observations at a time, and emergent constraints are no different in this respect. Robust constraint of a high-dimensional system requires large numbers of combined constraints so that the multiple compensating dimensions of uncertainty can be reduced (Reddington et al., 2017). We are reasonably confident that extension of our approach to more and varied observations will enable the uncertainty in aerosol radiative forcing to be reduced significantly.

Acknowledgements

This research was funded by the Natural Environment Research Council (NERC) under Grants NE/J024252/1 (GASSP), NE/I020059/1 (ACID-PRUF) and NE/P013406/1 (A-CURE); the European Union ACTRIS-2 project under grant 262254; the National Centre for Atmospheric Science (Yoshioka, Carslaw); and by the UK–China Research and Innovation Partnership Fund through the Met Office Climate Science for Service Partnership (CSSP) China as part of the Newton Fund. We made use of the N8 HPC facility funded from the N8 consortium and an Engineering and Physical Sciences Research Council Grant to use ARCHER (EP/K000225/1) and the JASMIN facility (www.jasmin.ac.uk/) via the Centre for Environmental Data Analysis funded by NERC and the UK Space Agency and delivered by the Science and Technology Facilities Council. We acknowledge the following additional funding: the Royal Society Wolfson Merit Award (Carslaw); a doctoral training grant from the Natural Environment Research Council and a CASE studentship with the Met Office Hadley Centre (Regayre).

References

Allen, R. J., Norris, J. R. and Wild, M.: Evaluation of multidecadal variability in CMIP5 surface solar radiation and inferred underestimation of aerosol direct effects over Europe, China, Japan, and India, J. Geophys. Res. Atmos., 118(12), 6311–6336, doi:10.1002/jgrd.50426, 2013.

Andreae, M. O., Jones, C. D. and Cox, P. M.: Strong present-day aerosol cooling implies a hot future., Nature, 435(7046), 1187–90, doi:10.1038/nature03671, 2005.

Andres, R. J. and Kasgnoc, A. D.: A time-averaged inventory of subaerial volcanic sulfur emissions, J. Geophys. Res. Atmos., 103(D19), 25251–25261, doi:10.1029/98JD02091, 1998.

Andrianakis, I., Vernon, I., McCreesh, N., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., Goldstein, M. and White, R. G.: History matching of a complex epidemiological model of human immunodeficiency virus transmission by using variance emulation, J. R. Stat. Soc. Ser. C Appl. Stat., 66(4), 717–740, doi:10.1111/rssc.12198, 2017.

Ban-weiss, G. A., Jin, L., Bauer, S. E., Bennartz, R., Liu, X., Zhang, K., Ming, Y., Guo, H. and Jiang, J. H.: Evaluating clouds, aerosols, and their interactions in three global climate models using satellite simulators and observations, J. Geophys. Res., 119, 10876–10901, doi:10.1002/2014JD021722, 2014.

Bellouin, N., Quaas, J., Morcrette, J. J. and Boucher, O.: Estimates of aerosol radiative forcing from the MACC re-analysis, Atmos. Chem. Phys., 13(4), 2045–2062, doi:10.5194/acp-13-2045-2013, 2013.

Boucher, O., Randall, D., Artaxo, P., Bretherton, C., Feingold, G., Forster, P., Kerminen, V.-M., Kondo, Y., Liao, H., Lohmann, U., Rasch, P., Satheesh, S. K., Sherwood, S., Stevens, B. and Zhang, X. Y.: Clouds and Aerosols, in Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by V. B. and P. M. M. Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. 571., 2013.

Boutle, I. A., Abel, S. J., Hill, P. G. and Morcrette, C. J.: Spatial variability of liquid cloud and rain: observations and microphysical effects, Q. J. R. Meteorol. Soc., 140(679), 583–594, doi:10.1002/qj.2140, 2014.

Carslaw, K. S., Lee, L. A., Reddington, C. L., Pringle, K. J., Rap, A., Forster, P. M., Mann, G. W., Spracklen, D. V, Woodhouse, M. T., Regayre, L. A. and Pierce, J. R.: Large contribution of natural aerosols to uncertainty in indirect forcing., Nature, 503(7474), 67–71, doi:10.1038/nature12674, 2013.

Carslaw, K. S., Gordon, H., Hamilton, D. S., Johnson, J. S., Regayre, L. A., Yoshioka, M. and Pringle, K. J.: Aerosols in the Pre-industrial Atmosphere, Curr. Clim. Chang. Reports, 3(1), 1–15, doi:10.1007/s40641-017-0061-2, 2017.

Cherian, R., Quaas, J., Salzmann, M. and Wild, M.: Pollution trends over Europe constrain global aerosol forcing as simulated by climate models, Geophys. Res. Lett., 41(6), 2176–2181, doi:10.1002/2013GL058715, 2014.

Craig, P. S., Goldstein, M., Seheult, A. H. and Smith, J. A.: Pressure Matching for Hydrocarbon Reservoirs: A Case Study in the Use of Bayes Linear Strategies for Large Computer Experiments, in Case Studies in Bayesian Statistics. Lecture

Notes in Statistics, vol 121., edited by Gatsonis, C., Hodges, J. S., Kass, R. E., McCulloch, R., Rossi, P., and Singpurwalla N.D., Springer, New York, NY., 1997.

Edwards, N. R., Cameron, D. and Rougier, J.: Precalibrating an intermediate complexity climate model, Clim. Dyn., 37(7–8), 1469–1482, doi:10.1007/s00382-010-0921-0, 2011.

Ghan, S., Wang, M., Zhang, S., Ferrachat, S., Gettelman, A., Griesfeller, J., Kipling, Z., Lohmann, U., Morrison, H., Neubauer, D., Partridge, D. G., Stier, P., Takemura, T., Wang, H. and Zhang, K.: Challenges in constraining anthropogenic aerosol effects on cloud radiative forcing using present-day spatiotemporal variability., Proc. Natl. Acad. Sci. U. S. A., 113(21), 5804–11, doi:10.1073/pnas.1514036113, 2016.

Ghan, S. J.: Technical note: Estimating aerosol effects on cloud radiative forcing, Atmos. Chem. Phys., 13(19), 9971–9974, doi:10.5194/acp-13-9971-2013, 2013.

Ghan, S. J. and Schwartz, S. E.: Aerosol Properties and Processes: A Path from Field and Laboratory Measurements to Global Climate Models, Bull. Am. Meteorol. Soc., 88(7), 1059–1083, doi:10.1175/BAMS-88-7-1059, 2007.

Grandey, B. S., Stier, P. and Wagner, T. M.: Investigating relationships between aerosol optical depth and cloud fraction using satellite, aerosol reanalysis and general circulation model data, Atmos. Chem. Phys., 13, 3177–3184, doi:10.5194/acp-13-3177-2013, 2013.

Gryspeerdt, E. and Stier, P.: Regime-based analysis of aerosol-cloud interactions, Geophys. Res. Lett., 39, 1–5, doi:10.1029/2012GL053221, 2012.

Gryspeerdt, E., Quaas, J. and Bellouin, N.: Constraining the aerosol influence on cloud fraction, J. Geophys. Res., 121, 3566–3583, doi:10.1002/2015JD023744.Abstract, 2016.

Gryspeerdt, E., Quaas, J., Ferrachat, S., Gettelman, A., Ghan, S., Lohmann, U., Morrison, H., Neubauer, D., Partridge, D. G., Stier, P., Takemura, T., Wang, H., Wang, M. and Zhang, K.: Constraining the instantaneous aerosol influence on cloud albedo., Proc. Natl. Acad. Sci. U. S. A., 114(19), 4899–4904, doi:10.1073/pnas.1617765114, 2017.

HadGEM3: HadGEM3: Met Office climate prediction model: HadGEM3 family, [online] Available from: http://www.metoffice.gov.uk/research/modelling–systems/unified–%0A962 model/climate–models/hadgem3. Accessed: March 2017, 2017. (Accessed 1 March 2017), 2017.

Halmer, M. M., Schmincke, H.-U. and H.-F., G.: The annual volcanic gas input into the atmosphere, in particular into the stratosphere: a global data set for the past 100 years, J. Volcanol. Geotherm. Res., 115, 511–528, 2002.

Hamilton, D. S., Lee, L. A., Pringle, K. J., Reddington, C. L., Spracklen, D. V and Carslaw, K. S.: Occurrence of pristine aerosol environments on a polluted planet., Proc. Natl. Acad. Sci. U. S. A., 111(52), 18466–71, doi:10.1073/pnas.1415440111, 2014.

Hartmann, D. L., Klein Tank, A. M. G., Rusticucci, M., Alexander, L. V., Brönnimann, S., Charabi, Y. A. R., Dentener, F. J., Dlugokencky, E. J., Easterling, D. R., Kaplan, A., Soden, B. J., Thorne, P. W., Wild, M. and Zhai, P.: Observations: Atmosphere and surface, Clim. Chang. 2013 Phys. Sci. Basis Work. Gr. I Contrib. to Fifth Assess. Rep. Intergov. Panel Clim. Chang., 9781107057, 159–254, doi:10.1017/CBO9781107415324.008, 2013.

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, Bull. Am. Meteorol. Soc., BAMS-D-15-00135.1, doi:10.1175/BAMS-D-15-00135.1, 2016.

Kahn, R. A.: Reducing the Uncertainties in Direct Aerosol Radiative Forcing, Surv. Geophys., 33(3–4), 701–721, doi:10.1007/s10712-011-9153-z, 2012.

Kooperman, G. J., Pritchard, M. S., Ghan, S. J., Wang, M., Somerville, R. C. J. and Russell, L. M.: Constraining the influence of natural variability to improve estimates of global aerosol indirect effects in a nudged version of the Community Atmosphere Model 5, J. Geophys. Res. Atmos., 117(D23), n/a-n/a, doi:10.1029/2012JD018588, 2012.

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, Atmos. Chem. Phys., 10(15), 7017–7039, doi:10.5194/acp-10-7017-2010, 2010.

Lebo, Z. J. and Feingold, G.: On the relationship between responses in cloud water and precipitation to changes in aerosol, Atmos. Chem. Phys., 14, 11817–11831, doi:10.5194/acp-14-11817-2014, 2014.

Lebsock, M., Morrison, H. and Gettelman, A.: Microphysical implications of cloud-precipitation covariance derived from satellite remote sensing, J. Geophys. Res. Atmos., 118(12), 6521–6533, doi:10.1002/jgrd.50347, 2013.

Lee, L. A., Carslaw, K. S., Pringle, K. J., Mann, G. W. and Spracklen, D. V.: Emulation of a complex global aerosol model to quantify sensitivity to uncertain parameters, Atmos. Chem. Phys., 11(23), 12253–12273, doi:10.5194/acp-11-12253-2011, 2011.

Lee, L. A., Pringle, K. J., Reddington, C. L., Mann, G. W., Stier, P., Spracklen, D. V., Pierce, J. R. and Carslaw, K. S.: The magnitude and causes of uncertainty in global model simulations of cloud condensation nuclei, Atmos. Chem. Phys., 13(17), 8879–8914, doi:10.5194/acp-13-8879-2013, 2013.

Lee, L. A., Reddington, C. L. and Carslaw, K. S.: On the relationship between aerosol model uncertainty and radiative forcing uncertainty., Proc. Natl. Acad. Sci. U. S. A., 113(21), 5820–7, doi:10.1073/pnas.1507050113, 2016.

Leibensperger, E. M., Mickley, L. J., Jacob, D. J., Chen, W. T., Seinfeld, J. H., Nenes, A., Adams, P. J., Streets, D. G., Kumar, N. and Rind, D.: Climatic effects of 1950-2050 changes in US anthropogenic aerosols-Part 1: Aerosol trends and radiative forcing, Atmos. Chem. Phys., 12(7), 3333–3348, doi:10.5194/acp-12-3333-2012, 2012.

Li, J., Han, Z. and Xie, Z.: Model analysis of long-term trends of aerosol concentrations and direct radiative forcings over East Asia, Tellus B Chem. Phys. Meteorol., 65(1), 20410, doi:10.3402/tellusb.v65i0.20410, 2013.

Liepert, B. and Tegen, I.: Multidecadal solar radiation trends in the United States and Germany and direct tropospheric aerosol forcing, J. Geophys. Res., 107(D12), doi:10.1029/2001JD000760, 2002.

Liu, Y. and Gupta, H. V.: Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, Water Resour. Res., 43(7), 1–18, doi:10.1029/2006WR005756, 2007.

Mann, G. W., Carslaw, K. S., Spracklen, D. V., Ridley, D. A., Manktelow, P. T., Chipperfield, M. P., Pickering, S. J. and Johnson, C. E.: Description and evaluation of GLOMAP-mode: a modal global aerosol microphysics model for the UKCA composition-climate model, Geosci. Model Dev., 3(2), 519–551, doi:10.5194/gmd-3-519-2010, 2010.

Mann, G. W., Carslaw, K. S., Reddington, C. L., Pringle, K. J., Schulz, M., Asmi, a., Spracklen, D. V., Ridley, D. a., Woodhouse, M. T., Lee, L. a., Zhang, K., Ghan, S. J., Easter, R. C., Liu, X., Stier, P., Lee, Y. H., Adams, P. J., Tost, H., Lelieveld, J., Bauer, S. E., Tsigaridis, K., van Noije, T. P. C., Strunk, a., Vignati, E., Bellouin, N., Dalvi, M., Johnson, C. E., Bergman, T., Kokkola, H., von Salzen, K., Yu, F., Luo, G., Petzold, a., Heintzenberg, J., Clarke, a., Ogren, J. a., Gras, J., Baltensperger, U., Kaminski, U., Jennings, S. G., O'Dowd, C. D., Harrison, R. M., Beddows, D. C. S., Kulmala, M., Viisanen, Y., Ulevicius, V., Mihalopoulos, N., Zdimal, V., Fiebig, M., Hansson, H.-C., Swietlicki, E. and Henzing, J. S.: Intercomparison and evaluation of global aerosol microphysical properties among AeroCom models of a range of complexity, Atmos. Chem. Phys., 14(9), 4679–4713, doi:10.5194/acp-14-4679-2014, 2014.

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., Haak, H., Jungclaus, J., Klocke, D., Matei, D., Mikolajewicz, U., Notz, D., Pincus, R., Schmidt, H. and Tomassini, L.: Tuning the climate of a global model, J. Adv. Model. Earth Syst., 4(8), 1–18, doi:10.1029/2012MS000154, 2012.

McCoy, D. T., Tan, I., Hartmann, D. L., Zelinka, M. D. and Storelvmo, T.: On the relationships among cloud cover, mixed-phase partitioning, and planetary albedo in GCMs, J. Adv. Model. Earth Syst., 8(2), 650–668, doi:10.1002/2015MS000589, 2016.

McNeall, D., Williams, J., Booth, B., Betts, R., Challenor, P., Wiltshire, A. and Sexton, D.: The impact of structural error on parameter constraint in a climate model, Earth Syst. Dyn., 7(4), 917–935, doi:10.5194/esd-7-917-2016, 2016.

Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M. and Stainforth, D. A.: Quantification of modelling uncertainties in a large ensemble of climate change simulations, Nature, 430(August 2004), 768–772, doi:10.1038/nature02771, 2004.

Myhre, G., Berglen, T. F., Johnsrud, M., Hoyle, C. R., Berntsen, T. K., Christopher, S. a., Fahey, D. W., Isaksen, I. S. a., Jones, T. a., Kahn, R. a., Loeb, N., Quinn, P., Remer, L., Schwarz, J. P. and Yttri, K. E.: Modelled radiative forcing of the direct aerosol effect with multi-observation evaluation, Atmos. Chem. Phys., 9(4), 1365–1392, doi:10.5194/acp-9-1365-2009, 2009.

Myhre, G., Shindell, D., Bréon, F., Collins, W., Fuglestvedt, J., Huang, J., Koch, D., Lamarque, J.-F., Lee, D., Mendoza, B, Nakajima, T., Robock, A. and Stephens, G.: Anthropogenic and natural radiative forcing, in Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, edited by V. B. and P. M. M. Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. 659., 2013.

O'Connor, F. M., Johnson, C. E., Morgenstern, O., Abraham, N. L., Braesicke, P., Dalvi, M., Folberth, G. A., Sanderson, M.

G., Telford, P. J., Voulgarakis, A., Young, P. J., Zeng, G., Collins, W. J. and Pyle, J. A.: Evaluation of the new UKCA climate-composition model-Part 2: The troposphere, Geosci. Model Dev., 7(1), 41–91, doi:10.5194/gmd-7-41-2014, 2014.

Penner, J. E., Xu, L. and Wang, M.: Satellite methods underestimate indirect climate forcing by aerosols, Proc. Natl. Acad. Sci. U. S. A., 108(33), 13404–13408, doi: 10.1073/pnas.1018526108, 2011.

Petters, M. D. and Kreidenweis, S. M.: A single parameter representation of hygroscopic growth and cloud condensation nucleus activity, Atmos. Chem. Phys., 7(8), 1961–1971, doi:10.5194/acp-7-1961-2007, 2007.

Pujol, G., Iooss, B. and Janon, A., with contributions from: Boumhaout, K., Da Veiga, S., Delage, T., Fruth, J., Gilquin, L., Guillaume, J., Le Gratiet, L., Lemaitre, P., Nelson, B. L., Monari, F., Oomen, R., Ramos, B., Roustant, O., Song, E., Staum, J., Touati, T. and Weber, F.: Sensitivity: Global Sensitivity Analysis of Model Outputs, R package version 1.14.0, 2017.

Quaas, J., Ming, Y., Menon, S., Takemura, T., Wang, M., Penner, J. E., Gettelman, A. and Lohmann, U.: Aerosol indirect effects – general circulation model intercomparison and evaluation with satellite data, Atmos. Chem. Phys., 9, 8697–8717, 2009.

Quaas, J., Stevens, B., Stier, P., Lohmann, U. and Physics, P.: Interpreting the cloud cover – aerosol optical depth relationship found in satellite data using a general circulation model, Atmos. Chem. Phys., 10, 6129–6135, doi:10.5194/acp-10-6129-2010, 2010.

Reddington, C. L., Carslaw, K. S., Stier, P., Schutgens, N., Coe, H., Liu, D., Allan, J., Browse, J., Pringle, K. J., Lee, L. A., Yoshioka, M., Johnson, J. S., Regayre, L. A., Spracklen, D. V., Mann, G. W., Clarke, A., Hermann, M., Henning, S., Wex, H., Kristensen, T. B., Leaitch, W. R., Pöschl, U., Rose, D., Andreae, M. O., Schmale, J., Kondo, Y., Oshima, N., Schwarz, J. P., Nenes, A., Anderson, B., Roberts, G. C., Snider, J. R., Leck, C., Quinn, P. K., Chi, X., Ding, A., Jimenez, J. L., Zhang, Q., Reddington, C. L., Carslaw, K. S., Stier, P., Schutgens, N., Coe, H., Liu, D., Allan, J., Browse, J., Pringle, K. J., Lee, L. A., Yoshioka, M., Johnson, J. S., Regayre, L. A., Spracklen, D. V., Mann, G. W., Clarke, A., Hermann, M., Henning, S., Wex, H., Kristensen, T. B., Leaitch, W. R., Pöschl, U., Rose, D., Andreae, M. O., Schmale, J., Kondo, Y., Oshima, N., Schwarz, J. P., Nenes, A., Anderson, B., Roberts, G. C., Snider, J. R., Leck, C., Quinn, P. K., Chi, X., Ding, A., Jimenez, J. L. and Zhang, Q.: The Global Aerosol Synthesis and Science Project (GASSP): Measurements and Modeling to Reduce Uncertainty, Bull. Am. Meteorol. Soc., 98(9), 1857–1877, doi:10.1175/BAMS-D-15-00317.1, 2017.

Regayre, L., Johnson, J. S., Yoshioka, M., Pringle, K. J., Sexton, D. M. H., Booth, B. B. B., Lee, L. A., Bellouin, N. and Carslaw, K. S.: Aerosol and host climate model parameters are both important sources of uncertainty in aerosol ERF, Atmos. Chem. Phys., 18, 9975-10006, doi:10.5194/acp-18-9975-2018, 2018.

Regayre, L. A., Pringle, K. J., Booth, B. B. B., Lee, L. A., Mann, G. W., Browse, J., Woodhouse, M. T., Rap, A., Reddington, C. L. and Carslaw, K. S.: Uncertainty in the magnitude of aerosol-cloud radiative forcing over recent decades, Geophys. Res. Lett., 41, 9040–9049, doi:10.1002/2014GL062029, 2014.

Regayre, L. A., Pringle, K. J., Lee, L. A., Rap, A., Browse, J., Mann, G. W., Reddington, C. L., Carslaw, K. S., Booth, B. B. B. and Woodhouse, M. T.: The Climatic Importance of Uncertainties in Regional Aerosol–Cloud Radiative Forcings over Recent Decades, J. Clim., 28(17), 6589–6607, doi:10.1175/JCLI-D-15-0127.1, 2015.

Ridley, D. A., Heald, C. L., Ridley, K. J. and Kroll, J. H.: Causes and consequences of decreasing atmospheric organic aerosol in the United States., Proc. Natl. Acad. Sci. U. S. A., 115(2), 290–295, doi:10.1073/pnas.1700387115, 2018.

Rodrigues, L. F. S., Vernon, I. and Bower, R.: Constraints on galaxy formation models from the galaxy stellar mass function and its evolution, Mon. Not. R. Astron. Soc., 466, 2418–2435, doi:10.1093/mnras/stw3269, 2017.

Saltelli, A., Tarantola, S. and Chan, K. P.-S.: A Quantitative Model-Independent Method for Global Sensitivity Analysis of Model Output, Technometrics, 41(1), 39–56, doi:10.1080/00401706.1999.10485594, 1999.

Salter, J. M. and Williamson, D.: A comparison of statistical emulation methodologies for multi-wave calibration of environmental models, Environmetrics, 27(8), 507–523, doi:10.1002/env.2405, 2016.

Samset, B., Myhre, G. and Schulz, M.: Upward adjustment needed for aerosol radiative forcing uncertainty, Nat. Clim. Chang., 4(April), 13–15, doi:10.1038/nclimate2170, 2014.

Schutgens, N. A. J., Partridge, D. G. and Stier, P.: The importance of temporal collocation for the evaluation of aerosol models with observations, Atmos. Chem. Phys., 16(2), 1065–1079, doi:10.5194/acp-16-1065-2016, 2016a.

Schutgens, N. A. J., Gryspeerdt, E., Weigum, N., Tsyro, S., Goto, D., Schulz, M. and Stier, P.: Will a perfect model agree with perfect observations? The impact of spatial sampling, Atmos. Chem. Phys., 16(10), 6335–6353, doi:10.5194/acp-16-6335-2016, 2016b.

Seinfeld, J. H., Bretherton, C., Carslaw, K. S., Coe, H., DeMott, P. J., Dunlea, E. J., Feingold, G., Ghan, S., Guenther, A. B., Kahn, R., Kraucunas, I., Kreidenweis, S. M., Molina, M. J., Nenes, A., Penner, J. E., Prather, K. A., Ramanathan, V., Ramaswamy, V., Rasch, P. J., Ravishankara, A. R., Rosenfeld, D., Stephens, G. and Wood, R.: Improving our fundamental understanding of the role of aerosol-cloud interactions in the climate system., Proc. Natl. Acad. Sci. U. S. A., 113(21), 5781–90, doi:10.1073/pnas.1514043113, 2016.

Sexton, D. M. H., Murphy, J. M., Collins, M. and Webb, M. J.: Multivariate probabilistic projections using imperfect climate models part I: outline of methodology, Clim. Dyn., 38(11–12), 2513–2542, doi:10.1007/s00382-011-1208-9, 2011.

Sexton, D. M. H., Karmalkar, A., Murphy, J. and Booth, B. B. B. .: The elicitation of distributions of parameters in HadGEM3 versions GA4 and GA7 for use in perturbed parameter ensembles, Hadley Cent. Tech. note, Met Off. U.K., 2017.

Shindell, D. T., Lamarque, J. F., Schulz, M., Flanner, M., Jiao, C., Chin, M., Young, P. J., Lee, Y. H., Rotstayn, L., Mahowald, N., Milly, G., Faluvegi, G., Balkanski, Y., Collins, W. J., Conley, A. J., Dalsoren, S., Easter, R., Ghan, S., Horowitz, L., Liu, X., Myhre, G., Nagashima, T., Naik, V., Rumbold, S. T., Skeie, R., Sudo, K., Szopa, S., Takemura, T., Voulgarakis, A., Yoon, J. H. and Lo, F.: Radiative forcing in the ACCMIP historical and future climate simulations, Atmos. Chem. Phys., 13(6), 2939–2974, doi:10.5194/acp-13-2939-2013, 2013.

Stier, P.: Limitations of passive satellite remote sensing to constrain global cloud condensation nuclei, Atmos. Chem. Phys., 16, 6595-6607, doi:10.5194/acp-16-6595-2016, 2016.

Terai, C. R., Wood, R. and Kubar, T. L.: Satellite estimates of precipitation susceptibility in low-level marine stratiform clouds, J. Geophys. Res., 120, 8878–8889, doi:10.1002/2015JD023319.Received, 2015.

Turnock, S. T., Spracklen, D. V., Carslaw, K. S., Mann, G. W., Woodhouse, M. T., Forster, P. M., Haywood, J., Johnson, C. E., Dalvi, M., Bellouin, N. and Sanchez-Lorenzo, A.: Modelled and observed changes in aerosols and surface solar radiation over Europe between 1960 and 2009, Atmos. Chem. Phys., 15(16), 9477–9500, doi:10.5194/acp-15-9477-2015, 2015.

van der Werf, G. R., Randerson, J. T., Giglio, L., Collatz, G. J., Mu, M., Kasibhatla, P. S., Morton, D. C., DeFries, R. S., Jin, Y. and van Leeuwen, T. T.: Global fire emissions and the contribution of deforestation, savanna, forest, agricultural, and peat fires (1997–2009), Atmos. Chem. Phys., 10(23), 11707–11735, doi:10.5194/acp-10-11707-2010, 2010.

West, R. E. L., Stier, P., Jones, A., Johnson, C. E., Mann, G. W., Bellouin, N., Partridge, D. G. and Kipling, Z.: The importance of vertical velocity variability for estimates of the indirect aerosol effects, Atmos. Chem. Phys., 14(12), 6369–6393, doi:10.5194/acp-14-6369-2014, 2014.

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L. and Yamazaki, K.: History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble, Clim. Dyn., 41(7–8), 1703–1729, doi:10.1007/s00382-013-1896-4, 2013.

Yi, B., Yang, P., Bowman, K. P. and Liu, X.: Aerosol-cloud-precipitation relationships from satellite observations and global climate model simulations, J. Appl. Remote Sens., 6(1), 63503, doi:http://dx.doi.org/10.1117/1.JRS.6.063503, 2012.

Yoshioka, M., Regayre, L., Pringle, K. J., Mann, G. W., Sexton, D. M. H., Johnson, C. E. and Carslaw, K. S.: Perturbed parameter ensembles of the HadGEM-UKCA composition-climate model to explore aerosol and radiative forcing uncertainty, J. Adv. Earth Syst., in-prep, 2018.

Zhang, S., Wang, M., Ghan, S. J., Ding, A., Wang, H., Zhang, K., Neubauer, D., Lohmann, U., Ferrachat, S., Takemura, T., Gettelman, A., Morrison, H., Lee, Y., Shindell, D. T., Partridge, D. G., Stier, P. and Kipling, Z.: On the characteristics of aerosol indirect effect based on dynamic regimes in global climate models, Atmos. Chem. Phys., 16, 2765–2783, doi:10.5194/acp-16-2765-2016, 2016.

Zhang, Y., Wang, K. and He, J.: Multi-year application of WRF-CAM5 over East Asia-Part II: Interannual variability, trend analysis, and aerosol indirect effects, Atmos. Environ., 165, 222–239, doi:10.1016/j.atmosenv.2017.06.029, 2017.