Deliverable D13.5: The value of measurements in the reduction in global model uncertainty

Ken Carslaw, Leighton Regayre, Jill Johnson (University of Leeds)

Work package no	**WP*13***
Deliverable no.	**D*13.5***
Lead beneficiary	
Deliverable type	☑ R (Document, report) □ DEC (Websites, patent fillings, videos, etc.) □ OTHER: please specify ....
Dissemination level	☑ PU (public) □ CO (confidential, only for members of the Consortium, incl Commission)
Estimated delivery date	**Month *41***
Actual delivery date	***01/07/2018***
Version	
Comments	

D13.5: THE VALUE OF MEASUREMENTS IN THE REDUCTION IN GLOBAL MODEL UNCERTAINTY

13.5.1 INTRODUCTION

This report describes how ACTRIS measurements of aerosol microphysical properties help to reduce the uncertainty in a global aerosol-climate model (HadGEM3-UKCA). We apply a rigorous statistical methodology in which the uncertainty in the model is calculated using around 1 million 'model variants' that sample 26 uncertainties in aerosol emissions and processes. We then quantify the reduction in model uncertainty achieved using only the model variants that produce plausible results when compared to the ACTRIS measurements. The model variants were generated from a perturbed parameter ensemble of the model, followed by model emulation, Monte Carlo sampling and history matching. We show the ACTRIS measurements enable a substantial reduction in model uncertainty.

13.5.2 METHODOLOGY

13.5.2.1 Sampling model uncertainty

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 in the HadGEM-UKCA climate model. These model variants (parameter combinations) define the 'prior' model uncertainty (which can be defined by a pdf of model output), which we then constrain by identifying variants that produce plausible outputs compared to ACTRIS observations. Model variants that produce results outside of the observational uncertainty range are considered implausible and are rejected. Likewise, the forcings that these model variants calculate are also rejected, which therefore enables a direct link to be established between the constraint of aerosol properties and the constraint of aerosol forcing.

We define observational constraint as finding the full set of model variants that are plausible when compared against observations. We can therefore estimate the prior (unconstrained) and remaining (observationally constrained) uncertainty range of the model. This approach is different to model tuning, which produces only one result on the right side of Fig 1 with no information about uncertainty. However, the process of adjusting the model to agree better is with observations is often misleadingly called constraint.

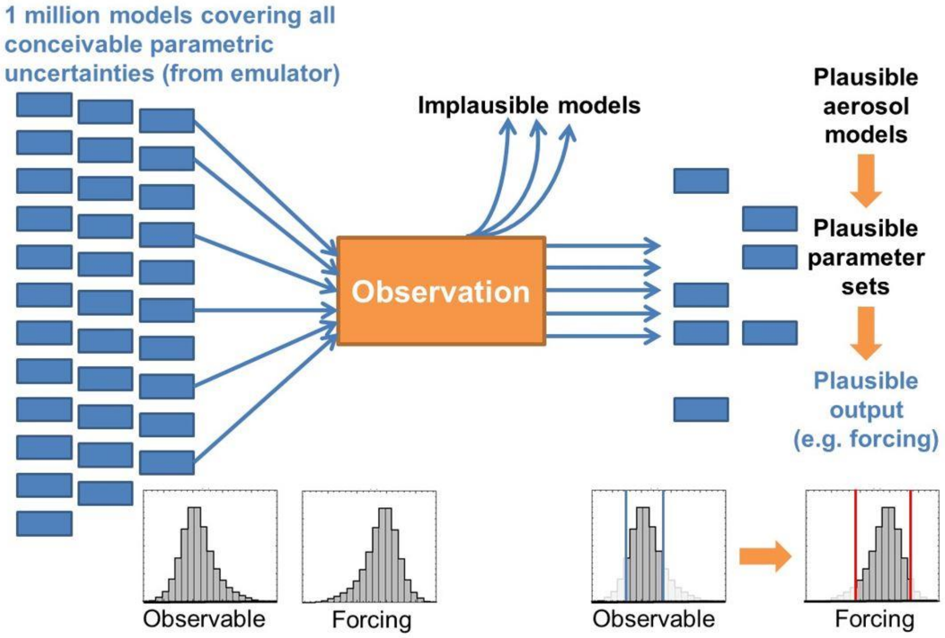

1 million models covering all conceivable parametric uncertainties (from emulator)
Implausible models
Observation
Plausible aerosol models
Plausible parameter sets
Plausible output (e.g. forcing)
Observable
Forcing
Observable
Forcing

Figure 1. Schematic of the methodology for observational constraint of parametric model uncertainty. From (Johnson *et al.*, 2018)

Most observational constraint studies are severely limited by the very small number of models used, which makes it impossible to reach robust statistical conclusions about model uncertainty (Carslaw *et al.*, 2018). In a multi-model ensemble such as in Aerocom the number of models is often about ten to twenty, 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). The large sample size allows us to robustly relate the uncertainty on the left side of Fig. 1 (in the form of a pdf) to the observationally constrained uncertainty on the right side.

The methodology is described in detail in three papers (Johnson *et al.*, 2018; Regayre *et al.*, 2018; Yoshioka *et al.*, 2018). In brief, the steps involved are (Figure 1):

1. A perturbed parameter ensemble (PPE) of the HadGEM3-UKCA aerosol-chemistry-climate model was created. In this ensemble combinations of 26 aerosol emissions and processes (parameters) were perturbed simultaneously. These causes of model uncertainty were identified as important for aerosols and aerosol-cloud forcing in our previous research (Carslaw *et al.*, 2013; Lee *et al.*, 2013; Regayre *et al.*, 2014, 2015). The PPE consists of two sets of 235 single-year simulations which differ only in the anthropogenic aerosol emissions prescribed (1850 and 2008). The experiment was designed to fill the 26-dimensional parameter space optimally using the 235 parameter combinations.

2. Emulators were built using data from the PPE (step 1). These emulators define (within quantifiable uncertainty) how aerosol properties and aerosol radiative forcing vary over the 26-dimensional parameter space. We validate each emulator's ability to reproduce model output, then use them to sample the 1 million Monte Carlo points from the parameter space to produce the set of model variants on the left side of Fig 1.

This step is essential because, with 26 dimensions of model uncertainty, the 235 PPE simulations are sparsely distributed. Filling the space more densely using output from global climate models is computationally expensive. However, emulators allow us to densely sample of the multi-dimensional parameter space and conduct robust statistical analyses.

3. We identify which of the 1 million model variants are consistent with the ACTRIS measurements within the uncertainty ranges of the individual measurements. This reduced set of variants defines the ways in which uncertain parameters can be combined to reproduce multiple observations and is equivalent to identifying thousands of equally plausible tuned HadGEM3-UKCA models. This procedure is often called 'history matching' or 'pre-calibration' (Craig *et al.*, 1997; Edwards, Cameron and Rougier, 2011; Williamson *et al.*, 2013; Lee, Reddington and Carslaw, 2016; Andrianakis *et al.*, 2017).

4. The reduction in aerosol radiative forcing uncertainty is quantified by comparing the uncertainty from the original sample of 1 million model variants with the uncertainty in the observationally plausible variants.

13.5.2.2 Constraint methodology

The constraint approach involves ruling out model variants (parameter combinations from the emulator) that are judged as implausible against measurements. We do this by calculating an *implausibility metric* (*I*) for each of the 1 million variants (*x;* (Williamson *et al.*, 2013; McNeall *et al.*, 2016; Andrianakis *et al.*, 2017)), which weights the difference between the model and observations by the uncertainties in both:

$$I(x) = \frac{|\, z - E[\eta(x)]\, |}{\sqrt{[Var\big(\phi(x)\big) + Var(\epsilon)]}}, \qquad (1)$$

where z is the measurement and $E[\eta(x)]$ is the estimate of model output calculated using the emulator $\eta(x)$. In the denominator $Var\big(\phi(x)\big)$ is the variance in the emulator prediction and $Var(\epsilon)$ is the variance in the measurement. In essence, if the confidence in the measurements and emulator is high (the denominator is small) then we can be confident that the difference between the model and measurement in the numerator is meaningful in terms of model skill, so a large model-measurement difference can be used to rule out a part of parameter space represented by the variant *x* (the implausibility is high). However, if the measurement uncertainty is large or the emulator is a poor representation of the model (has large uncertainty), then we cannot be confident that the variant is producing implausible results, so we retain that variant (the implausibility is small). The emulator uncertainty is known at all points in parameter space, so is used directly in the implausibility metric calculation.

$Var(\epsilon)$ has four additive components:

$$Var(\epsilon) = Var(\epsilon_{MEAS}) + Var(\epsilon_{IAV}) + Var(\epsilon_{SP}) + Var(\epsilon_{TEMP}) \qquad (2)$$

which are:

1. Measurement (instrument) uncertainty $\big(Var(\epsilon_{MEAS})\big)$
2. Inter-annual variability in aerosol properties $\big(Var(\epsilon_{IAV})\big)$ accounting for the fact that we may wish to match observations and the model for the correct calendar month but not for the correct year.

3. Spatial co-location uncertainty $\left(Var(\epsilon_{SP})\right)$ accounting for the potentially large spatial variability of point measurements below the grid scale of the model (Schutgens *et al.*, 2016, 2017).
4. Temporal co-location uncertainty component $\left(Var(\epsilon_{TEMP})\right)$ accounting for the fact that the temporal sampling of an observation may not match to the temporal sampling of the model (e.g. a ship track through the grid-box over a short time period which is compared with a monthly-mean model value (Schutgens *et al.*, 2017).

We call $\epsilon_{IAV} + \epsilon_{SP} + \epsilon_{TEMP}$ the model-measurement representation error, or *representation error* for short (Reddington *et al.*, 2017) because it defines the error associated with how well the measured aerosol property is represented in the model.

A challenge with this semi-automatic constraint procedure, as with any model-measurement comparison, is that we cannot a priori account for model structural error: i.e., the model-measurement error may be very large because the model lacks essential processes and therefore we should not expect the measured values to lie within the range produced by sampling the model parameters. To account for this possibility we include a filtering step in which we examine the mean implausibility (across the 1 million model variants) for each measurement in each month and decide whether very high values may indicate structural errors. These measurements are then removed from the analysis and flagged for future investigation. We also allow a defined number (or percentage) of observations (tolerance, *T*) to exceed a defined implausibility threshold (θ). For example, we might rule out a model variant if it has implausibility metrics larger than θ=3 for more than T =20% of the observations (i.e., bias is 3 times the expected error). Values of *T* and θ are subjective, and were chosen for each observational constraint according to the relative effect on the aerosol properties being constrained.

For all aerosol properties we assume a measurement/instrument uncertainty of 10%, a spatial co-location uncertainty of 20%, and a temporal sampling uncertainty of 10% based on typical values (Schutgens *et al.*, 2016, 2017; Reddington *et al.*, 2017). The sampling uncertainties can vary across the different observed aerosol properties as well as spatially and temporally, but we have not attempted to account for these variations here. The inter-annual uncertainty was estimated based on a 30-year simulation and is shown for a range of aerosol properties in Figure 3.

In addition to the implausibility metric we use the normalized mean absolute error factor (NMAEF; (Gustafson and Yu, 2012) to quantitatively compare the average model output to measurements. The NMAEF metric is defined as:

$$NMAEF = \frac{\sum|\,\mathrm{A(i)}-\mathrm{B(i)}|}{\sum \mathrm{A(i)}}, \qquad (3)$$

where A is the observation and B is the model variant in the case where the model overestimates the measurement. If the model underestimates the measurement A is the model variant and B the observation.

The NMAEFs tells us the factor by which the mean unconstrained sample value over/under predicts the measurements. We multiply the NMAEF by -1 in cases where the mean model value underpredicts the measurement at each location/month combination on average so as to understand the general tendency for over- or under-prediction. Large absolute NMAEF values suggest substantial over/underprediction.

However, the NMAEFs alone do not provide a comprehensive insight into the degree of model-measurement agreement. For example, large positive NMAEF values can result from comparing near-zero measurement values to relatively low model values. Relative to comparisons in other locations and/or months, the large NMAEF case may actually provide a reasonable model-observation comparison. Furthermore, the NMAEFs do not account for the sources of uncertainty important for model/measurement comparison outlined above. Hence, we interpret the NMAEFs alongside the implausibility metrics.

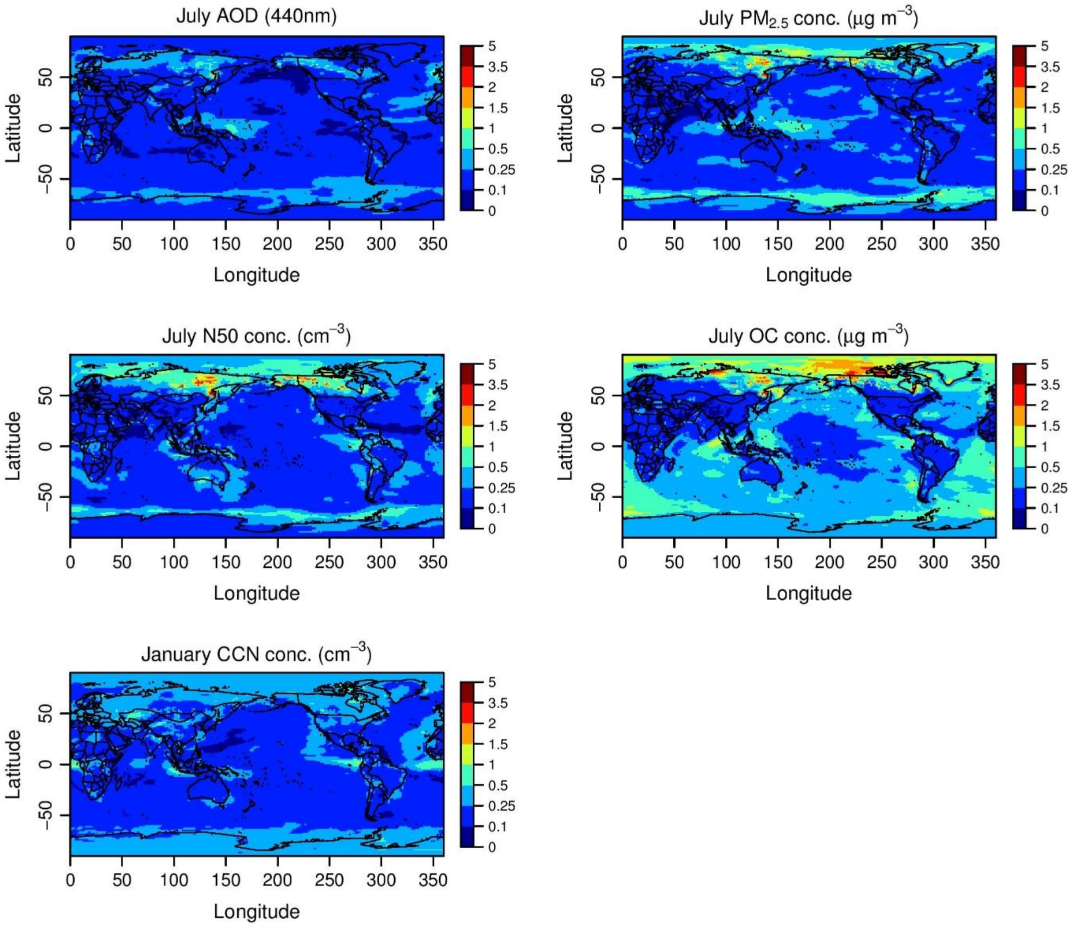

July AOD (440nm)
July PM$_{2.5}$ conc. (µg m^{-3})
July N50 conc. (cm^{-3})
July OC conc. (µg m^{-3})
January CCN conc. (cm^{-3})
Latitude
Longitude
−50
0
50
0
50
100
150
200
250
300
350
5
3.5
2
1.5
1
0.5
0.25
0.1
0

Figure 2. The relative standard deviation of July-mean aerosol properties over a 30-year period, used in the estimation of the inter-annual variability component of the implausibility measure, ϵ_{IAV}. This relative standard deviation was generated from the analysis of a UKCA hindcast simulation over the period of 1980-2009. Values refer to surface-level PM2.5, N50, OC and CCN.

13.5.2.3 Measurements

ACTRIS particle number size distribution measurements were used to calculate monthly mean concentrations of particles larger than 50 nm and 3 nm diameter (N50 and N3). Individual measurements with missing data in some bins were removed from the analysis. Averaging over multiple years allows us to use all of the remaining measurement data in the constraint process. Particle concentrations were averaged over multiple years (as available on the EBAS website) within each month at each location. In cases where two or more sites are within the same model gridbox the data from each site was given equal weighting when calculating the multi-year monthly mean values. We account for inter-annual variability in the calculation of our implausibility metrics (Figure 2). The ACTRIS measurement used for model constraint are summarized in table 1.

Index	Station name	Latitude site	Latitude model	Longitude site	Longitude model	Filtered dataset months used
1	BEO Moussala	42.17	42.5	23.58	22.5	Jan - Dec
2	Schauinsland and Jungfraujoch	47.91 and 46.55	47.5	7.91 and 7.99	7.5	Jan – Dec
3	Hohenpeissenberg and Schneefernerhaus	47.8 and 47.42	47.5	11.01 and 10.98	11.25	Jan – Dec
4	Melpitz	51.53	52.5	12.93	11.25	Jan – Dec
5	Hyytiala	61.85	62.5	24.28	11.25	Jan – Dec
6	Pallas (Sammaltunturi)	68.0	67.5	24.14	22.5	Jan – Dec
7	Gual Pahari	28.42	27.5	77.15	78.75	Jan – Dec
8	Mt Cimone	44.18	45	10.7	11.25	Jan – Dec
9	Preila	55.38	55	21.03	22.5	Jan – Mar; Jul – Nov
10	Birkenes	58.38	57.5	8.25	7.5	Jan – Dec
11	Zeppelin Mountain (Ny-Alesund)	78.91	80	11.89	11.25	Jan – Dec
12	Troll	-72.01	-72.5	2.53	3.75	Jan – Dec
13	Aspvreten	58.8	60	17.38	18.75	Jan – Dec

Table 1: Number concentration measurement locations used in the analysis. Longitude and latitudes of sites, as well as the closest longitude and latitude at the resolution of our model data are provided. Finally we indicate the months for which data is available after filtering missing values.

13.5.3 RESULTS

13.5.3.1 Model-measurement comparison

In tables 2 and 3 we present the mean implausibility and NMAEF for N3 and N50 respectively at each of the ACTRIS measurement sites used in our analysis.

At several sites such as Hyytiala, Pallas, Gual Pahari, Mt Cimone and Aspvreten the mean N3 and N50 of the sample of model variants compares very well with observations across the year. At the Birkenes, Mt Cimone and Zeppelin Mountain sites there is excellent model-measurement agreement for much of the year. However, at these sites there are anomalous months where either N3 or N50 compares poorly (the mean implausibility is larger than around 3). In the case of Mt Cimone and Zeppelin Mountain in the final months of the year the comparison is so poor (the lower credible bound of the sample of implausibility metrics exceeds *I*=1) that the observations in these months are removed from the constraint process. These results suggest there is some structural deficiency in the model that needs to be addressed (missing particle nucleation events/processes) and/or some form of corruption in the data that was undetected by our data screening process.

The model-measurement N3 comparison is much better (lower mean implausibility) in winter than summer months at many of the sites (BEO Moussala, Schauinsland and Jungfraujoch, Hohenpeissenberg and Schneefernerhaus, Melpitz, Hyytiala, Mt Cimone, Birkenes and Aspvreten). However, there is little seasonality in the N50 implausibility metrics at these sites. Figure (Histograms) shows the pdfs of N3 and N50 for January and July from the unconstrained sample of one million model variants, as well as pdfs from the subsample constrained to match all measurements of N3 and N50. N3 concentrations in July in the unconstrained sample are far larger than N3 concentrations in January and N50 concentrations in both months. The seasonality in N3 implausibility metrics can partly be explained by the fact that the implausibility metrics scale with the magnitude of the measurements.

The NMAEFs are typically positive for both N3 and N50 at most sites, but are much larger for N3. These results suggest that in general the model over predicts both N3 and N50 measurements, although the model-measurement agreement is better for N50. Despite a tendency for over prediction of aerosol concentrations, the mean implausibility values are generally small (less than around 3; See Fig. 3). This suggests that the differences in model and measurement concentrations are relatively small compared to the variance terms in the denominator of equation (1).

For several winter-spring months at the Troll site the measured N3 and N50 concentrations do not agree with the range of modelled values and the measurements are removed from the constraint process. Troll is the only Southern Hemisphere measurement site used in this analysis. The model also compares poorly with N50 concentrations at this site, with the distribution of implausibility across the model variants being so large the observations are removed from the constraint process (by definition). In February (SH summer) the distribution of implausibility for Troll is tightly centered around a large mean implausibility – i.e., there is considerable disagreement between the measurements at this site and the majority of the model variants. In July the mean implausibility metric at Troll (2.78) is relatively high compared to other sites, but parts of parameter space are in reasonable agreement with the measurements (the distribution covers I=1, the cutoff value for using the measurement data). The consistently negative NMAEFs for N3 and N50 at Troll suggest that the model underpredicts aerosol concentrations in remote marine locations. Aerosol concentrations in remote locations are known to be affected by different aerosol emission, deposition and process parameters than densely populated Northern Hemisphere locations (Regayre *et al.*, 2015). If the model is structurally deficient in its capacity to simulate remote aerosol concentrations, it may be that a source of natural aerosols and/or aerosol growth processes are missing from the model.

Station	**J**	**F**	**M**	**A**	**M**	**J**	**J**	**A**	**S**	**O**	**N**	**D**
BEO Moussala	0.55 (5.6)	0.74 (6.5)	1.64 (10.9)	2.51 (10.2)	3.25 (14.7)	3.07 (12.1)	4.66 (16.1)	3.84 (10.5)	2.70 (10.8)	1.96 (14.4)	0.79 (12.6)	0.50 (5.9)
Schauinsland and Jungfraujoch	0.66 (2.9)	1.06 (4.2)	1.64 (4.6)	2.44 (4.7)	2.89 (5.7)	2.64 (3.6)	3.12 (4.2)	3.16 (4.6)	2.28 (3.8)	1.80 (3.3)	0.73 (1.8)	0.67 (2.6)
Hohenpeissenberg and Schneefernerhaus	0.64 (1.2)	1.16 (1.9)	1.53 (1.9)	2.17 (3.5)	3.26 (5.3)	2.41 (3.1)	2.72 (3.4)	2.96 (4.1)	2.42 (3.7)	2.15 (3.3)	0.88 (1.7)	0.79 (2.2)
Melpitz	0.99 (-1.2)	0.70 (-0.7)	0.64 (-0.5)	1.03 (1.0)	2.04 (2.0)	3.22 (2.4)	2.82 (2.1)	2.67 (1.9)	1.26 (0.9)	0.70 (-0.5)	0.72 (-0.8)	0.77 (-0.9)
Hyytiala	0.71 (-0.8)	1.92 (-1.8)	1.50 (-1.3)	0.84 (-0.5)	1.68 (0.8)	2.10 (1.9)	1.88 (1.6)	3.02 (-1.5)	0.91 (-0.4)	0.82 (-0.6)	0.69 (-0.7)	2.30 (-2.5)
Pallas (Sammaltunturi)	0.89 (1.1)	0.80 (0.7)	0.62 (0.6)	0.59 (-0.6)	0.75 (0.6)	1.02 (0.9)	0.79 (-0.5)	1.04 (-0.7)	0.85 (-0.7)	0.71 (-0.6)	0.71 (-0.6)	0.79 (1.0)
Gual Pahari	0.95 (0.6)	0.92 (-0.5)	0.91 (-0.5)	0.98 (-0.5)	1.14 (1.1)	1.32 (-0.7)	1.15 (-0.5)	1.14 (0.8)	1.14 (-0.5)	1.17 (-0.6)	1.09 (-0.6)	0.94 (-0.5)
Mt Cimone	0.66 (1.1)	0.87 (2.3)	1.34 (2.5)	2.28 (4.4)	2.71 (5.4)	2.36 (4.2)	2.60 (4.8)	2.74 (7.2)	2.30 (4.3)	1.77 (5.1)	4.91 (-6.2)	5.91 (-9.9)
Preila	0.87 (-1.5)	0.96 (-1.7)	0.64 (-0.8)	NA	NA	NA	1.78 (1.2)	2.03 (1.4)	1.41 (1.9)	1.40 (-1.4)	1.92 (-3.2)	NA
Birkenes	0.35 (0.8)	0.33 (1.1)	0.37 (0.5)	0.76 (1.6)	2.48 (8.0)	3.56 (14.4)	2.70 (7.3)	2.63 (4.9)	1.35 (2.2)	0.65 (0.8)	0.33 (1.0)	0.39 (0.8)
Zeppelin Mountain (Ny-Alesund)	1.19 (0.6)	0.82 (0.5)	1.12 (-0.5)	0.70 (-0.5)	0.67 (-0.3)	0.72 (-0.6)	1.01 (-0.9)	0.94 (-0.6)	0.45 (0.4)	0.98 (0.9)	0.82 (0.5)	7.70 (-25.2)
Troll	3.45 (-1.9)	5.15 (-3.8)	5.39 (-3.5)	5.01 (-3.3)	3.78 (-1.8)	2.51 (-1.0)	2.61 (-1.3)	1.78 (-0.9)	2.50 (-1.5)	4.29 (-2.3)	3.37 (-1.6)	1.19 (-0.6)
Aspvreten	0.70 (-0.7)	0.73 (-0.6)	0.78 (-0.5)	0.95 (0.7)	1.31 (1.0)	2.19 (2.1)	1.21 (1.0)	1.23 (0.9)	1.03 (0.5)	0.86 (0.5)	0.75 (-0.6)	0.68 (0.7)

Table 2: Unconstrained sample mean implausibility metrics and NMAEFs (in brackets) for N3. Implausibility metrics are shaded red for cases where the lower credible bound of implausibility metrics in the unconstrained sample is larger than I=1.

Station	J	F	M	A	M	J	J	A	S	O	N	D
BEO Moussala	3.13 (4.0)	2.98 (4.8)	3.00 (2.9)	2.17 (1.1)	2.72 (2.0)	2.82 (2.3)	2.37 (1.2)	1.66 (0.7)	2.09 (1.3)	2.91 (3.5)	3.16 (7.4)	2.88 (3.5)
Schauinsland and Jungfraujoch	2.52 (3.6)	2.95 (4.6)	2.61 (2.5)	2.66 (1.4)	3.06 (2.2)	2.59 (1.3)	2.92 (1.6)	3.00 (1.7)	2.77 (1.7)	2.91 (2.2)	2.36 (2.1)	2.78 (3.1)
Hohenpeissenberg and Schneefernerhaus	2.33 (1.7)	2.35 (1.6)	1.61 (0.7)	2.15 (0.9)	2.53 (1.3)	2.46 (1.3)	2.35 (1.3)	2.76 (1.6)	2.64 (1.5)	2.73 (1.7)	2.56 (2.0)	3.17 (2.8)
Melpitz	1.77 (-0.6)	1.76 (0.6)	1.68 (0.5)	1.87 (0.8)	2.18 (0.9)	2.28 (0.8)	2.32 (0.8)	2.03 (0.7)	1.69 (0.6)	1.78 (0.7)	2.13 (0.8)	1.98 (0.7)
Hyytiala	1.27 (0.7)	1.57 (-0.6)	1.91 (-0.6)	1.56 (0.5)	1.50 (0.5)	1.48 (0.6)	1.13 (0.5)	0.95 (0.4)	1.22 (0.4)	1.39 (0.5)	1.47 (0.6)	1.54 (0.8)
Pallas (Sammaltunturi)	1.63 (1.4)	1.41 (0.7)	1.48 (0.7)	1.41 (-0.5)	1.24 (-0.5)	0.99 (0.4)	0.65 (-0.3)	1.34 (-0.7)	0.91 (-0.4)	1.02 (0.5)	1.57 (1.0)	1.68 (1.8)
Gual Pahari	1.49 (0.5)	1.40 (0.6)	1.38 (0.6)	1.49 (0.7)	1.73 (0.8)	1.72 (0.6)	1.78 (-0.5)	1.70 (0.6)	1.72 (0.7)	1.68 (0.6)	1.47 (0.5)	1.53 (0.5)
Mt Cimone	1.83 (1.0)	2.40 (2.3)	2.19 (1.4)	2.37 (1.2)	2.66 (1.6)	2.57 (1.7)	2.12 (1.4)	2.64 (2.5)	1.95 (0.9)	2.85 (2.9)	2.63 (-0.9)	3.93 (-1.6)
Preila	3.38 (-1.8)	4.20 (-2.2)	2.56 (-1.1)	NA	NA	NA	2.44 (-0.8)	1.40 (0.6)	1.06 (0.4)	3.28 (-1.5)	4.38 (-2.4)	NA
Birkenes	1.43 (-0.6)	1.29 (0.7)	2.62 (-0.9)	1.63 (0.8)	2.74 (1.9)	2.39 (1.3)	2.21 (1.3)	0.98 (0.3)	0.88 (-0.3)	2.62 (-1.0)	1.41 (0.9)	1.43 (-0.6)
Zeppelin Mountain (Ny-Alesund)	1.41 (0.7)	0.88 (0.5)	1.42 (-0.6)	1.50 (-0.6)	1.11 (-0.4)	2.27 (-1.3)	1.98 (-1.3)	0.98 (-0.5)	0.88 (-0.5)	1.24 (1.1)	0.87 (0.5)	3.99 (-2.8)
Troll	6.64 (-5.7)	6.98 (-7.6)	6.44 (-4.9)	5.62 (-3.3)	4.23 (-1.8)	3.44 (-1.1)	2.78 (-1.0)	2.85 (-1.2)	3.00 (-1.0)	4.73 (-2.5)	6.11 (-4.1)	6.65 (-6.4)
Aspvreten	1.22 (0.6)	1.45 (-0.6)	1.49 (-0.6)	1.18 (0.4)	1.07 (0.4)	1.77 (0.7)	0.94 (0.4)	0.85 (0.3)	0.99 (0.3)	1.26 (-0.4)	1.27 (0.7)	1.54 (1.0)

Table 3: Unconstrained sample mean implausibility metrics and NMAEFs (in brackets) for N50. Implausibility metrics are shaded red for cases where the lower credible bound of implausibility metrics in the unconstrained sample is larger than I=1.

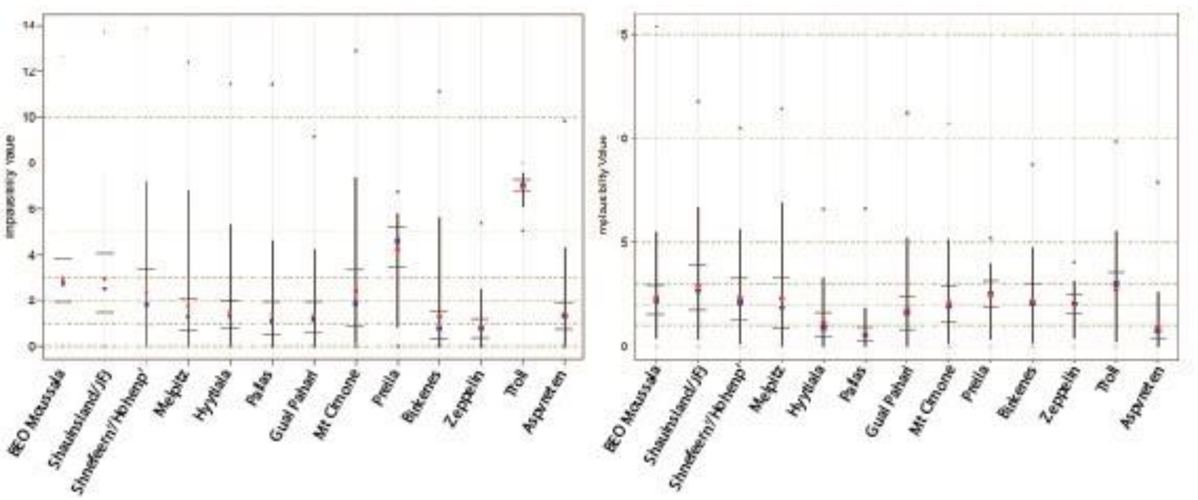


Figure 3. Implausibility metric for N50 for a) February and b) July at all 13 locations. For each observation location the range of the distribution across the model variants is shown by the outer crosses, the black bar corresponds to the 95% credible interval (2.5% to 97.5% empirical quantiles) and the small horizontal black lines within the bar show the inter-quartile range. The red circle corresponds to the mean and the blue square is the median.

ACTRIS (www.actris.eu) is supported by the European Commission under the Horizon 2020 – Research and Innovation Framework Programme, H2020-INFRAIA-2014-2015, Grant Agreement number: 654109

13.5.3.2 Constraint of model uncertainty

Figure 4 shows the unconstrained (prior) and constrained distributions of N3 and N50 averaged across the measurement locations in Europe. The constraint rules out large parts of the prior range of the model variants. When we constrain the model using individual months of data the number of model variants is reduced by about 50%. When we constrain using all months of data the number of variants is reduced by around 85%.

For N50, the mean over the European measurement sites (1, 2, 3, 4, 5, 6, 8, 9, 10 and 13 in table 1) is 1954 cm^{-3}. The model unconstrained mean is 3235 cm^{-3} (averaged across the sites and model variants) but after constraint this reduces to 2355 cm^{-3} – i.e., a reduction in the mean difference across the measurement sites by around 70%. The observed mean N50 in January is 1160 cm^{-3}. The model unconstrained mean is 1648 cm^{-3} and after constraint this reduces to 1063 cm^{-3}, which is very close to the measurements and a reduction of 80% in the model-measurement bias.

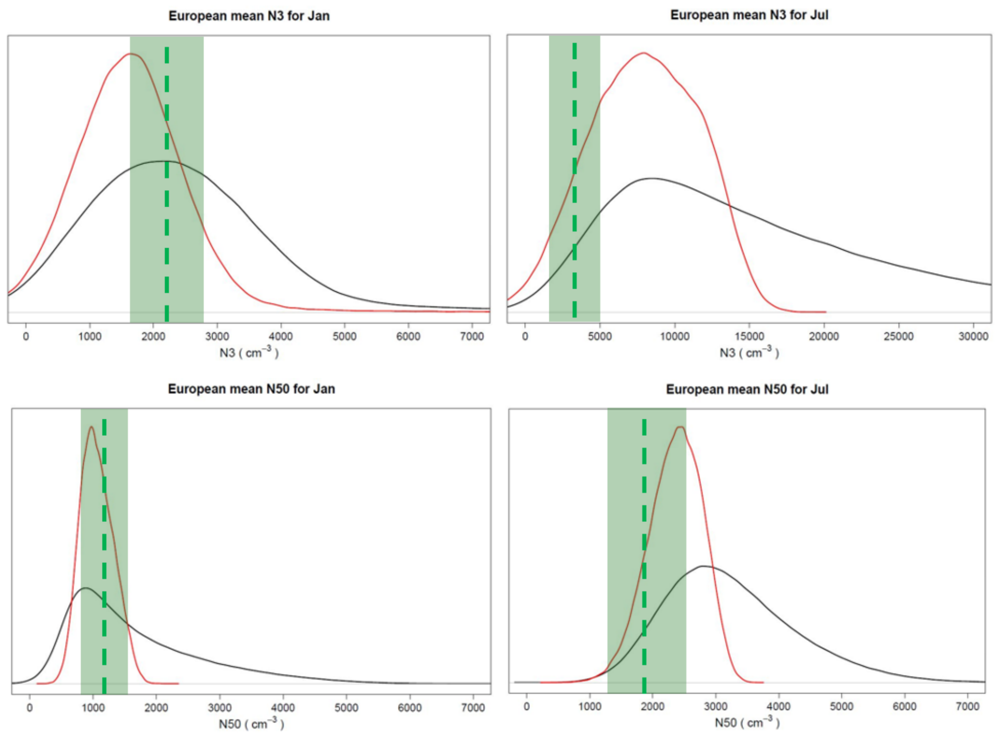


Figure 4. Prior and constrained model uncertainty distributions for N3 and N50 calculated across the European measurement locations. The black lines show the range of N3 and N50 of the 1 million model variants sampling the 26 model uncertainties. The red line shows the distribution after observational constraint. The green line and shading shows the measurement mean and approximate uncertainty, which we assume is dominated by the representation error (see section 13.5.2.2).

Figure 5 shows the 95% confidence range of the unconstrained and constrained samples of model variants. Over Europe the range is reduced substantially from about 50,000 cm^{-3} to 10,000 cm^{-3} for N3 and from about 10,000 cm^{-3} to 3,000 cm^{-3} for N50. The ACTRIS measurements therefore strongly constrain our model uncertainty.

Figure 6 shows that mean concentrations of N3 and N50 are substantially reduced across Europe, as expected from the probability distributions in Figure 4. The means concentrations are in much closer agreement with the measured values.

We also see that constraint using the small number of ACTRIS measurements helps to constrain the model uncertainty in other regions of the globe (Figure 7). We see reductions in the 95% confidence range for N50 mostly in polluted regions like N America and China, but also in marine regions affected by pollution outflow from these regions. The reductions in uncertainty, as well as the constrained values, would need to be evaluated against measurements in these locations. However, we expect many of the model uncertainties to be common to different regions (Lee, Reddington and Carslaw, 2016).

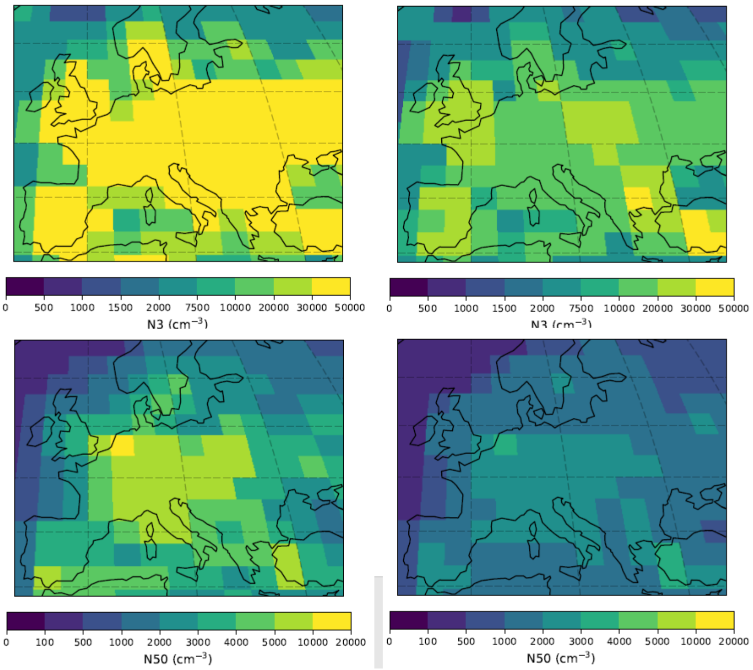


Figure 5. 95% confidence range of the unconstrained (left) and constrained samples of model variants for N3 (top) and N50 (bottom).

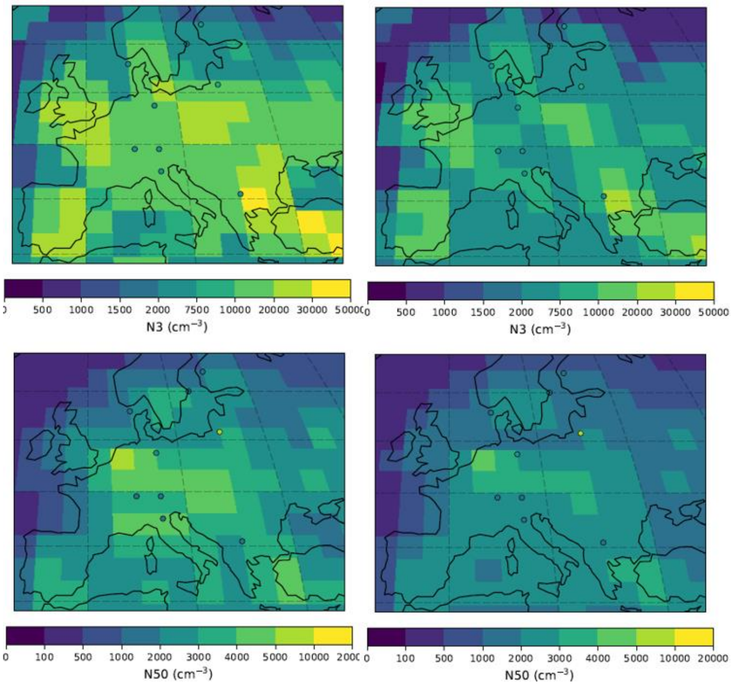


Figure 6. Mean N3 (top) and N50 (bottom) for the unconstrained (left) and constrained (right) samples of model variants. Measured particle concentrations are shown as dots.

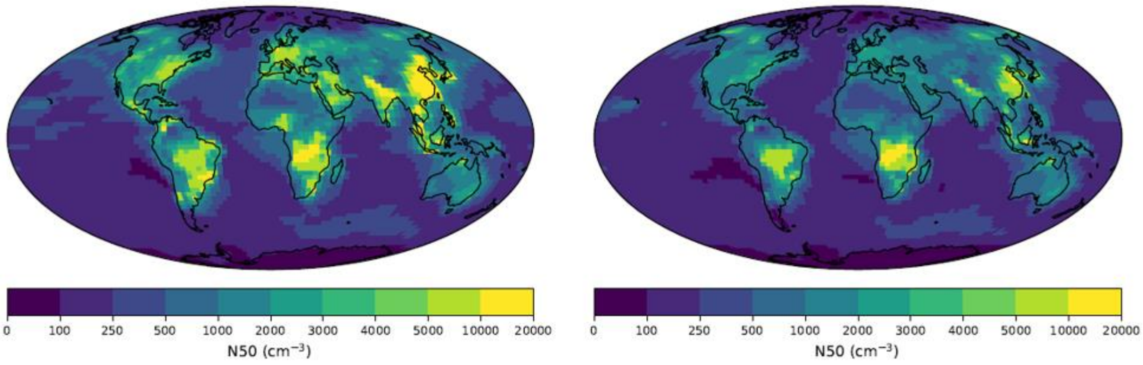


Figure 7. 95% confidence range of the unconstrained (left) and constrained samples of model variants for N50.

13.5.3.3 Constraint of model parameter ranges

Figure 8 shows the constraint of the prior parameter ranges using the N3 measurements in January or July. There is a large difference in the parameters that are constrained because of different parameter sensitivities in different seasons. In winter there is only very weak constraint of any parameters. The emission of primary particulate sulfate is weakly constrained: the likelihood of the smallest emission diameters and largest emission fluxes of these sub-grid particles is reduced a small amount. These reductions suggests that the prior ranges were producing too many particles, so this part of parameter space has been ruled out as implausible.

In summer, N3 measurements constrain a much wider range of parameters than in the winter. We see strong constraint of boundary layer nucleation rates (to the lower part of the prior range), the assumed pH of cloud water (to the upper range), anthropogenic SO_2 emissions (lower range), BVOC emissions (upper range), dry deposition velocities of both Aitken and accumulation mode aerosol (lower range), and the assumed hygroscopicity of organic material (upper range). Some of these constraints are consistent with some prior model variants producing too many particles. Nucleation rates are reduced and sink terms of small particles are increased, such as deposition rates of particles that increase the condensation sink. High pH is also consistent with sulphate production by reaction with ozone, and the sulphate also acts as a sink for nucleated particles.

Figure 9 compares the constraint caused by N3 and N50 over all months. For N50 we see very strong constraint of the diameter of primary carbonaceous and primary sulfate emitted particles, as well as the assumed diameter of the Aitken mode. Again, these constraints are consistent with the ruled-out parts of parameter space producing too many particles (larger particle diameters result in fewer emitted particles when the emitted mass flux is assumed constant). Boundary layer nucleation rates are also constrained to the lower part of the prior range, but the constraint is weaker than when using N3 measurements. This result (relatively weak constraint of nucleation rates but strong constraint of primary particle sizes) is fully consistent with the conclusion of earlier research (Reddington *et al.*, 2011), which showed that evidence for the role of nucleation in controlling N50 particles concentrations in the European boundary layer was not statistically robust when the uncertainty in particle emission diameters was taken into account.

Figure 10 shows the overall constraint of parameters by applying N3 and N50 measurements at ACTRIS sites over all months. The overall conclusion is that we can rule out large parts of parameter space that generate too many particles. We therefore see strong constraint of nucleation rates, primary particle diameters and deposition rates of particles that control the condensation sink of nucleating vapours and growing particles.

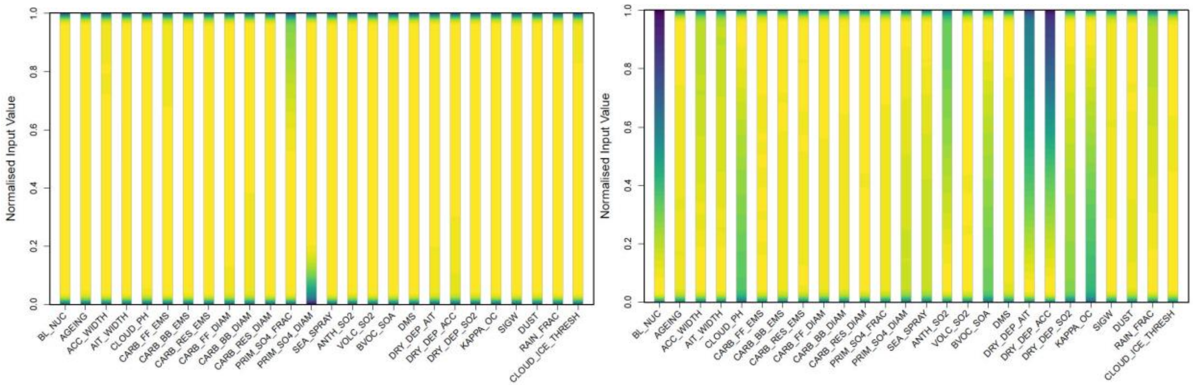
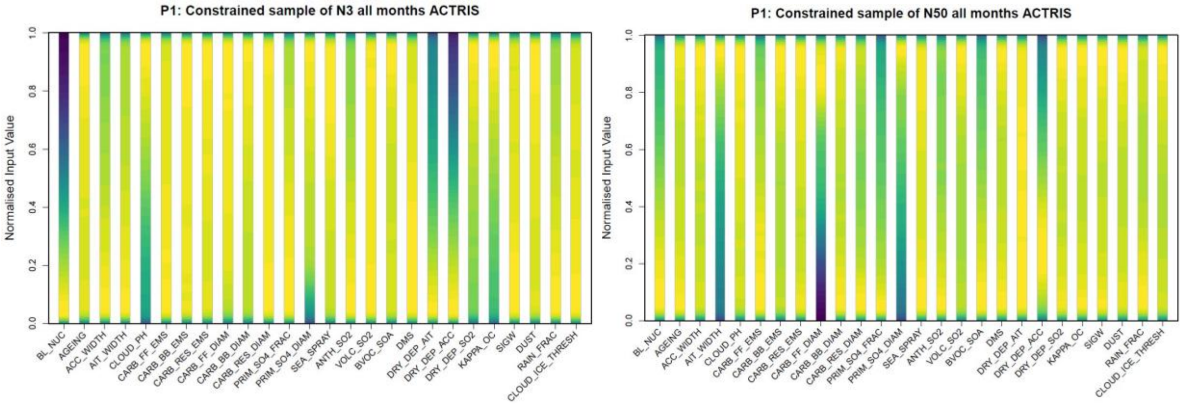


Figure 8. Constraint of model parameter ranges for two months (left, January; right, July) using N3 measurements only. The parameters are listed along the bottom of each figure. Each bar shows the normalised range of each parameter between 0 and 1 (see (Regayre *et al.*, 2018) for a definition of each parameter and its expert-elicited range. The colour of each bar indicates the probability (yellow is a probability of 1, i.e. no constraint of the prior assumption, and blue indicates a reduction in the probability of that part of parameter space in the constrained sample of variants.


P1: Constrained sample of N3 all months ACTRIS

P1: Constrained sample of N50 all months ACTRIS

Figure 9. Constraint of model parameter ranges using N3 measurements only (left) and N50 measurements only (right) for all months. The parameters are listed along the bottom of each figure. Each bar shows the normalised range of each parameter between 0 and 1 (see (Regayre *et al.*, 2018) for a definition of each parameter and its expert-elicited range. The colour of each bar indicates the probability (yellow is a probability of 1, i.e. no constraint of the prior assumption, and blue indicates a reduction in the probability of that part of parameter space in the constrained sample of variants.

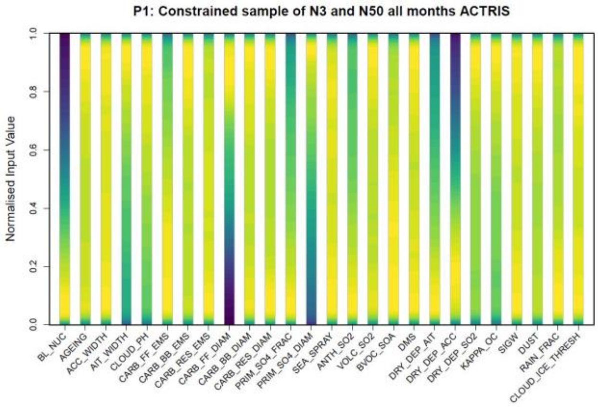


Figure 10. Constraint of model parameter ranges for using combined N3 and N50 measurements over all months. The parameters are listed along the bottom of each figure. Each bar shows the normalised range of each parameter between 0 and 1 (see (Regayre *et al.*, 2018) for a definition of each parameter and its expert-elicited range. The colour of each bar indicates the probability (yellow is a probability of 1, i.e. no constraint of the prior assumption, and blue indicates a reduction in the probability of that part of parameter space in the constrained sample of variants.

13.5.3.4 Constraint of simulated aerosol forcing

Our ultimate objective is to use aerosol measurements to reduce the uncertainty in model simulations of aerosol radiative forcing. Figure 11 shows that the aerosol radiative forcing over Europe and globally is only very weakly affected by the strong constraints on particle concentrations shown above. Over Europe there is a small shift on the most likely forcing to more negative values, but almost no change in the uncertainty. This effect is coming from a reduction in the number of particles in the constrained sample of variants combined with an increase in the probability of larger emitted primary particles. Clearly N3 and N50 measurements alone are insufficient to constrain all relevant aspects of the aerosol size distribution, so in further work we would extend the analysis to other size cutoffs. However, the weak constraint we see here is consistent with our study in which we included nine different types of measurements, including aerosol optical depth, PM and top-of-atmosphere fluxes (Johnson *et al.*, 2018).

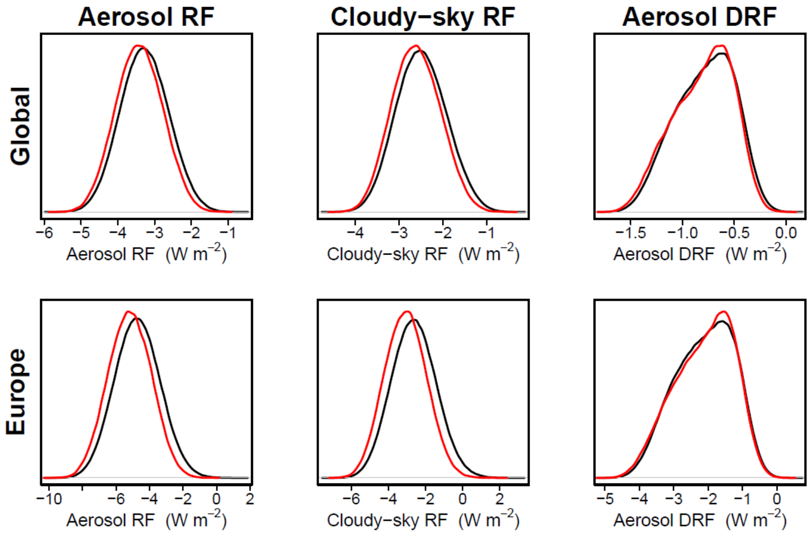


Figure 11. Unconstrained (black) and constrained (red) probability distributions of pre-industrial to present-day aerosol radiative forcing.

13.5.4 CONCLUSIONS

Our analysis of a large perturbed parameter ensemble shows that ACTRIS aerosol microphysics measurements provide a powerful constraint on some uncertain parameters in a global aerosol model. We have been able to rule out about 85% of our prior sampled parameter space using just measurements of N3 and N50 at 16 sites. Constraint leads to a reduction in model spread of about a factor of 3 for these simulated quantities and a strong sift in the mean distribution towards the measurements. These constraints also affect other (unmeasured) regions because we assume the parametric uncertainties are uniform globally,

The measurements are most effective at constraining model parameters related to particle formation rates, the size of emitted primary pollutant particles and deposition rates.

Constraint of aerosol forcing uncertainty is very weak using just these measurements. The strong constraints on N3 and N50 result in a small shift of the mean aerosol forcing over Europe to more negative values, but the uncertainty is not significantly affected. This result is consistent with our other work showing that even many more types of aerosol and in situ measurements are unlikely to be sufficient to constrain the forcings (Johnson *et al.*, 2018).

In order to constrain the forcing uncertainty further, it would be necessary to include measurements from a more diverse range of environments that represent both polluted and pristine (pre-industrial-like) conditions. In this study we used mainly ACTRIS measurements from Europe where the main sources of uncertainty (and hence the constrainable parameters) are related to anthropogenic emissions. We also propose to extend the analysis to more than N3 and N50 so that important changes in the size distribution can be constrained. We also argue that reduction in the recent decadal forcing may be easier than for the pre-industrial to present-day forcing because the parameters that control the recent forcing are more related to measurable quantities (Regayre *et al.*, 2014). In this regard, measurements of long-term trends would be highly valuable as a constraint on models, which is something that ACTRIS can contribute to.

Spatial representation error is a major source of uncertainty when constraining models. This important source of error may exceed the instrument uncertainty, so it should be a priority to characterize it for all measurement sites. This might be possible by conducting intensive campaign measurements around the ACTRIS sites, for example through carefully designed networks of well-characterised low-cost instruments on a routine basis.

13.5.5 REFERENCES

Andrianakis, I. *et al.* (2017) 'History matching of a complex epidemiological model of human immunodeficiency virus transmission by using variance emulation', *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 66(4), pp. 717–740. doi: 10.1111/rssc.12198.

Carslaw, K. S. *et al.* (2013) 'Large contribution of natural aerosols to uncertainty in indirect forcing.', *Nature*. Nature Publishing Group, 503(7474), pp. 67–71. doi: 10.1038/nature12674.

Carslaw, K. S. *et al.* (2018) 'Climate models are uncertain, but we can do something about it', *Eos, Transactions American Geophysical Union*, 99. doi: https://doi.org/10.1029/2018EO093757.

Craig, P. S. *et al.* (1997) 'Pressure Matching for Hydrocarbon Reservoirs: A Case Study in the Use of Bayes Linear Strategies for Large Computer Experiments', in C., G. et al. (eds) *Case Studies in*

Bayesian Statistics. Lecture Notes in Statistics, vol 121. 2nd editio. Springer, New York, NY.

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

Gustafson, W. I. and Yu, S. (2012) 'Generalized approach for using unbiased symmetric metrics with negative values: Normalized mean bias factor and normalized mean absolute error factor', *Atmospheric Science Letters*, 13(4), pp. 262–267. doi: 10.1002/asl.393.

Johnson, J. S. *et al.* (2018) 'The importance of comprehensive parameter sampling and multiple observations for robust constraint of aerosol radiative forcing', *Atmospheric Chemistry and Physics Discussions*, (March), pp. 1–38. doi: 10.5194/acp-2018-174.

Lee, L. A. *et al.* (2011) 'Emulation of a complex global aerosol model to quantify sensitivity to uncertain parameters', *Atmospheric Chemistry and Physics*, 11(23), pp. 12253–12273. doi: 10.5194/acp-11-12253-2011.

Lee, L. A. *et al.* (2013) 'The magnitude and causes of uncertainty in global model simulations of cloud condensation nuclei', *Atmospheric Chemistry and Physics*, 13(17), pp. 8879–8914. doi: 10.5194/acp-13-8879-2013.

Lee, L. A., Reddington, C. L. and Carslaw, K. S. (2016) 'On the relationship between aerosol model uncertainty and radiative forcing uncertainty.', *Proceedings of the National Academy of Sciences of the United States of America*, 113(21), pp. 5820–7. doi: 10.1073/pnas.1507050113.

McNeall, D. *et al.* (2016) 'The impact of structural error on parameter constraint in a climate model', *Earth System Dynamics*, 7(4), pp. 917–935. doi: 10.5194/esd-7-917-2016.

Reddington, C. L. *et al.* (2011) 'Primary versus secondary contributions to particle number concentrations in the European boundary layer', *Atmospheric Chemistry and Physics*, 11(23), pp. 12007–12036. doi: 10.5194/acp-11-12007-2011.

Reddington, C. L. *et al.* (2017) 'The Global Aerosol Synthesis and Science Project (GASSP): Measurements and Modeling to Reduce Uncertainty', *Bulletin of the American Meteorological Society*, 98(9), pp. 1857–1877. doi: 10.1175/BAMS-D-15-00317.1.

Regayre, L. *et al.* (2018) 'Aerosol and host climate model parameters are both important sources of uncertainty in aerosol ERF', *Atmospheric Chemistry and Physics*, submitted.

Regayre, L. A. *et al.* (2014) 'Uncertainty in the magnitude of aerosol-cloud radiative forcing over recent decades', *Geophysical Research Letters*, 41, pp. 9040–9049. doi: 10.1002/2014GL062029.

Regayre, L. A. *et al.* (2015) 'The Climatic Importance of Uncertainties in Regional Aerosol–Cloud Radiative Forcings over Recent Decades', *Journal of Climate*, 28(17), pp. 6589–6607. doi: 10.1175/JCLI-D-15-0127.1.

Schutgens, N. *et al.* (2017) 'On the spatio-temporal representativeness of observations', *Atmospheric Chemistry and Physics*, 17(16), pp. 9761–9780. doi: 10.5194/acp-17-9761-2017.

Schutgens, N. A. J. *et al.* (2016) 'Will a perfect model agree with perfect observations? The impact of spatial sampling', *Atmospheric Chemistry and Physics*, 16(10), pp. 6335–6353. doi: 10.5194/acp-16-6335-2016.

Williamson, D. *et al.* (2013) 'History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble', *Climate Dynamics*, 41(7–8), pp. 1703–1729. doi: 10.1007/s00382-013-1896-4.

Yoshioka, M. *et al.* (2018) 'Perturbed parameter ensembles of the HadGEM-UKCA composition-climate model to explore aerosol and radiative forcing uncertainty', *Journal of Advances in Earth Systems*, p. in-prep.