

Statistical inference for ensembles of models

Lindsay lee

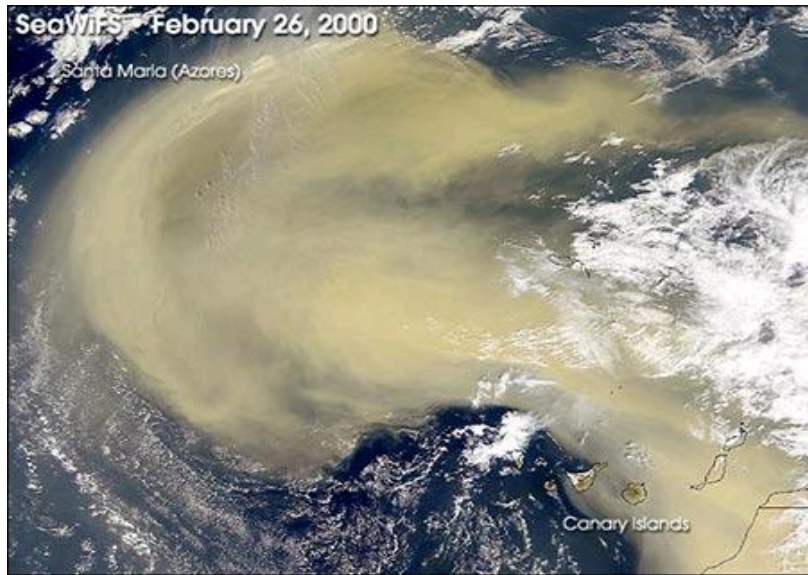
l.a.lee@leeds.ac.uk

Outline I

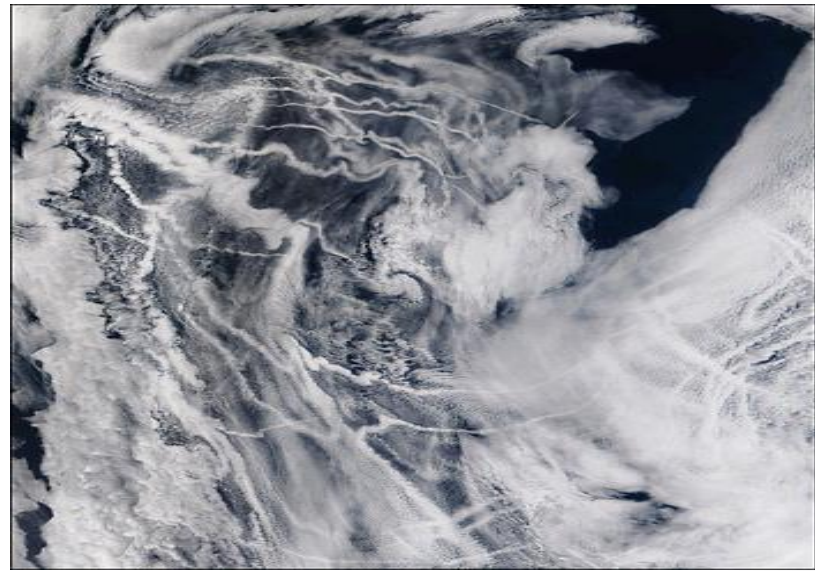
- Global aerosol and the climate
- Modelling global aerosol and model outputs
- Statistics for uncertainty
- Global aerosol modelling as a statistical problem
- Model ensembles
- Multi-model ensembles including discussion
- Perturbed physics ensembles including discussion
- History matching and calibration
- Using limited observations (maybe)

Global aerosol and climate

Direct effect



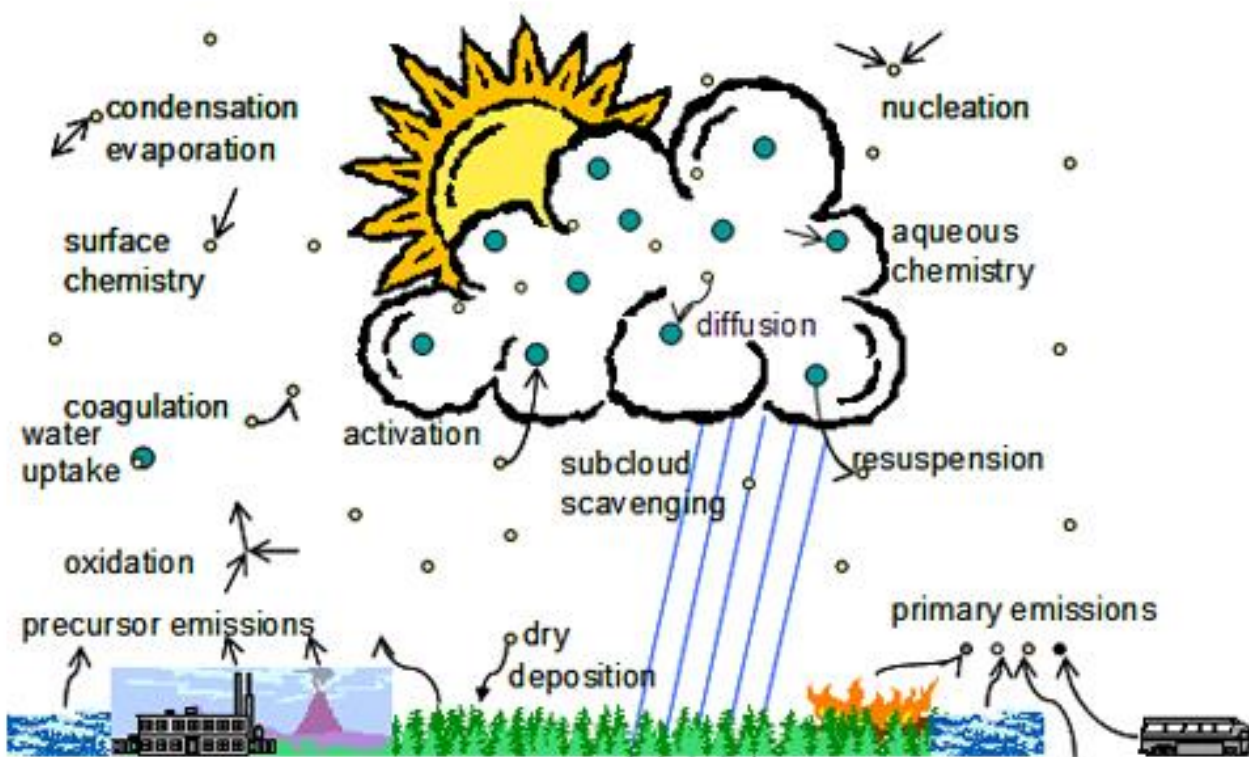
Indirect effect



Used for air quality too but our focus is climate relevant outputs

Modelling global aerosol

- We use the global aerosol model **GLOMAP** (Mann et al. 2010)
- A microphysical modal model simulating the evolution of global aerosol including sulphate, sea-salt, dust and black carbon

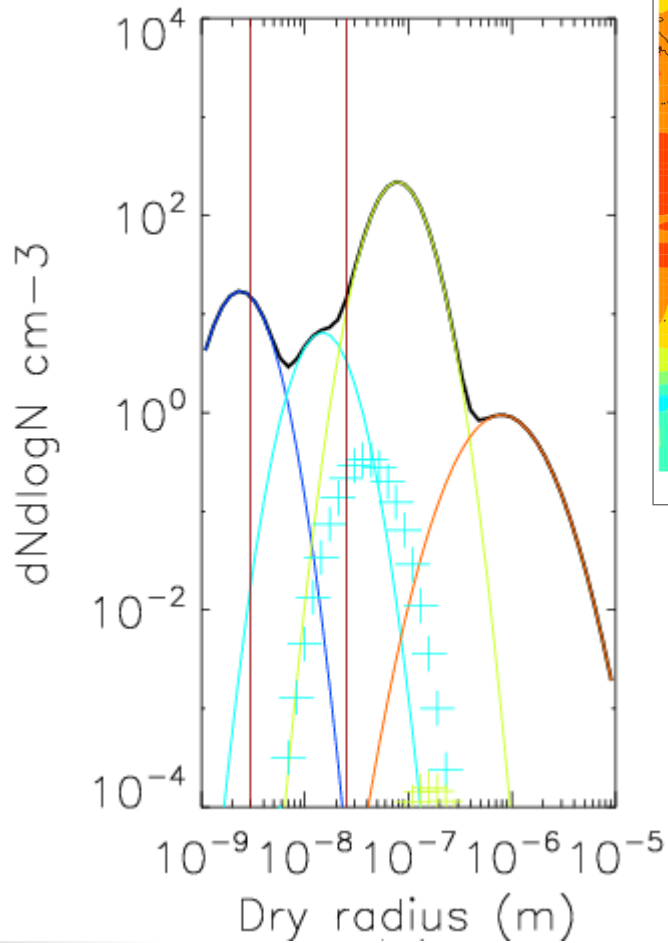


http://www.pnl.gov/atmospheric/research/aci/aci_aerosol_indeffects.stm

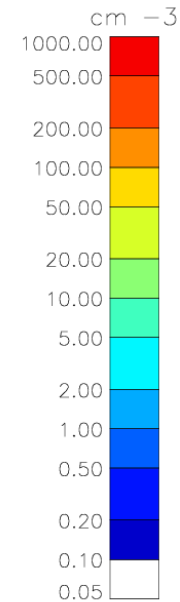
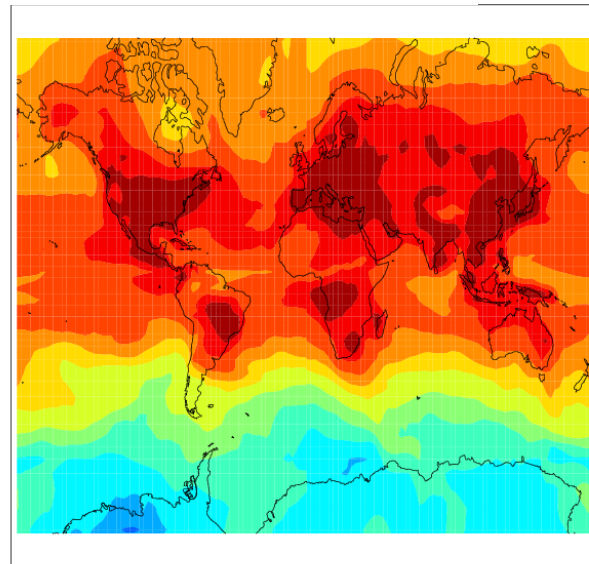
Aerosol model outputs

– from a single run

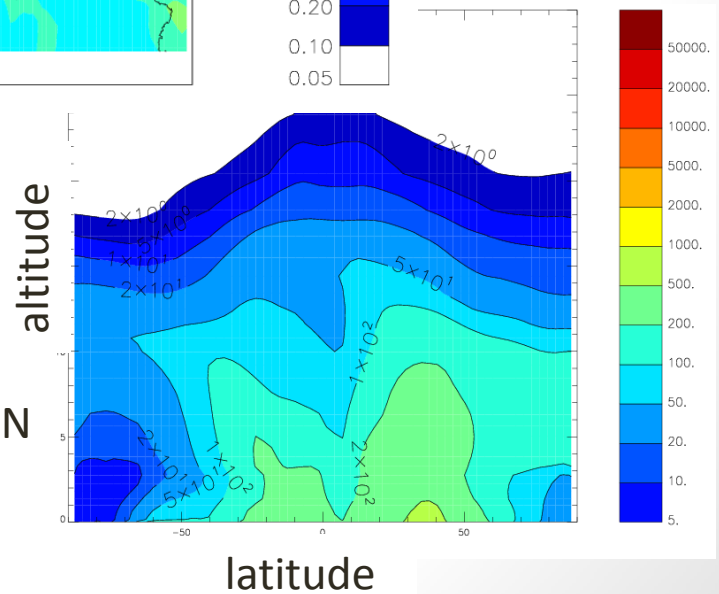
Pacific grid box



CCN concentration



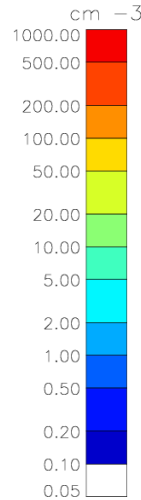
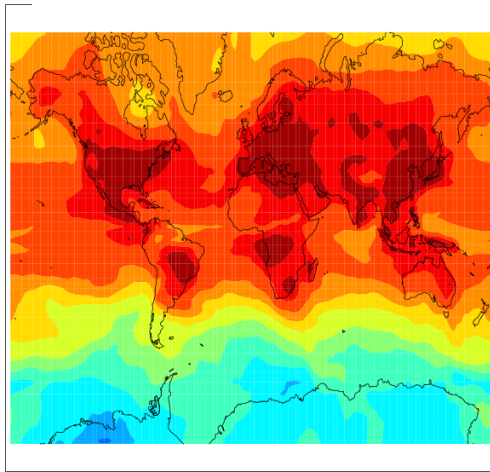
Zonal mean CCN concentration



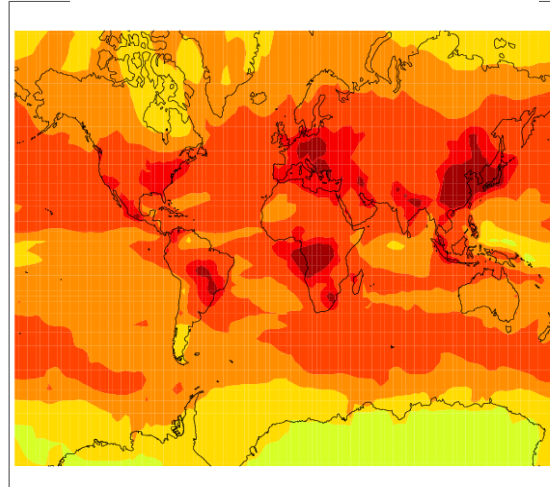
Aerosol model outputs

- from multiple runs with different model settings

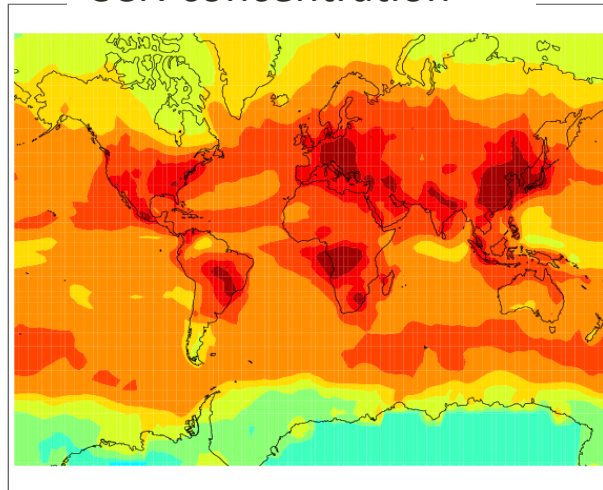
CCN concentration



CCN concentration



CCN concentration



Why use statistics for understanding uncertainty?

- Traceable (and testable) framework
- Using probability allows us to infer things that we couldn't otherwise.

Quantification

How reliable is my model prediction?

Understanding

Are my conclusions robust?
How can the uncertainty be reduced?



Global aerosol model notation

$$\mathbf{Y} = f(\mathbf{X})$$

- Aerosol model f
- Aerosol model inputs/drivers/initial conditions/structures \mathbf{X}
- Aerosol model output \mathbf{Y}

Bold letters are used to denote X and Y are vectors or matrices

- \mathbf{X} and \mathbf{Y} are random variables with probability distributions

$$\mathbf{X} \sim G_X \text{ and } \mathbf{Y} \sim G_Y$$

- f is also uncertain

Capital letters are used to denote X and Y are random variables (unknown)

Probability distributions for uncertainty

$$\mathbf{X} \sim G_{\mathbf{X}} \text{ and } \mathbf{Y} \sim G_{\mathbf{Y}}$$

- Properties of the probability density functions $G_{\mathbf{X}}$ and $G_{\mathbf{Y}}$ represent the uncertainty usually described by its moments
- Mean is the central value – **model estimate**
- Variance is the spread – **uncertainty about the model estimate**
- Skewness and kurtosis reveal more about the shape
- Our **interest lies in the mean and variance** but the shape is important for quantification and interpretation

Mean and variance

- $E[Y]$: estimated by $\hat{\mu} = \sum_{j=1}^n \frac{y_j}{n}$ - the expected value
- $Var[Y]$: estimated by $\hat{\sigma}^2 = \frac{\sum_{j=1}^n (y_j - \hat{\mu})^2}{n-1}$ $j = 1, 2, \dots, n$

Conditional probability

- **Always** working with conditional probabilities
- We are in fact looking for G_Y given G_X therefore we use $Y|X$ and $Y|X_i$ and $Y|X_{i,j}$
- There are also other factors upon which Y could be conditioned, e.g. **the model, the model runs, observations, initial conditions**
 - We usually use notation to show what we HAVE considered
 - Experience shows that this work opens up the discussion on dependency of results
- The **robustness** of Y can be tested by changing the conditions
- With GLOMAP we **use sampling** to find conditional probabilities conditioned on parameter uncertainty

Why do model ensembles?

- Model uncertainty:
 - Initial condition uncertainty – mostly weather models
 - Parametric uncertainty – all computer models
 - Structural uncertainty – different computer models
- Repetition necessary to measure uncertainty
 - G_X and G_Y can not be measured by a single number
 - We can't derive things analytically
- It's important to know **which uncertainties are represented** in an ensemble and which aren't.

Model ensembles

- Initial condition ensembles
 - Repeated model runs with changing initial condition
 - Common for weather prediction
- Multi-model ensembles (MME)
 - Compare models attempting to simulate the same thing
 - Target structural uncertainty
- Perturbed physics ensembles (PPE)
 - Repeated model runs with changing parameter values
 - Target parameter uncertainty
- GLOMAP is part of the international AEROCOM initiative – an MME
- In GLOMAP we do PPE to target parameter uncertainty where parameters represent process and emission uncertainties

Y is usually similar between the ensembles but X changes in each

Discussion: language

- I define the uncertainties according to what I think can be investigated in particular ensembles

What is the difference between a parameter perturbation and a structural perturbation?

Can boundary conditions/initial conditions/parameters/inputs be represented by parameters?

Multi-model ensembles - MMES

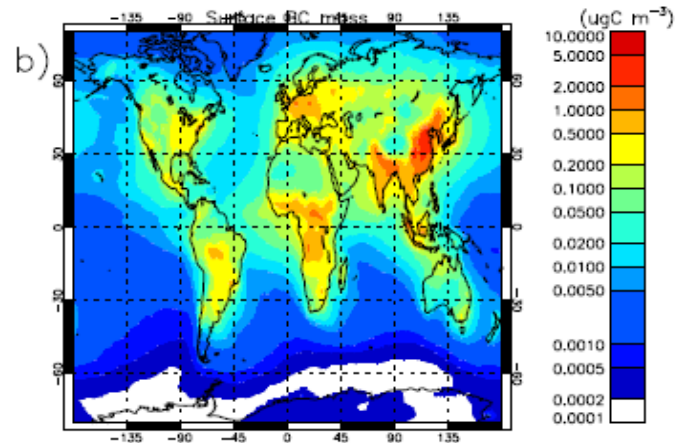
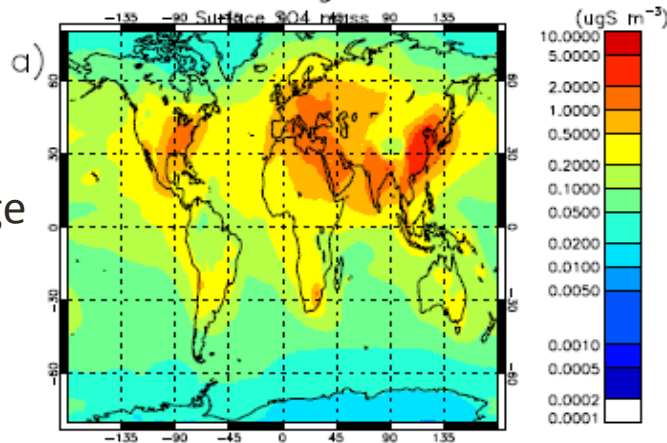
- Target structural uncertainty
- Look at models with different structures denoted \mathbf{X}
 - Processes are coded differently
 - Processes are treated differently
- Use summary statistics to represent G_Y
- How do the different ways processes are treated affect the model results?
- Use maps and graphics and comparison to observations to try to understand structural uncertainty $G_{\mathbf{X}}$ where \mathbf{X} here represents different model structures

AEROCOM – MME of aerosol models

G_Y

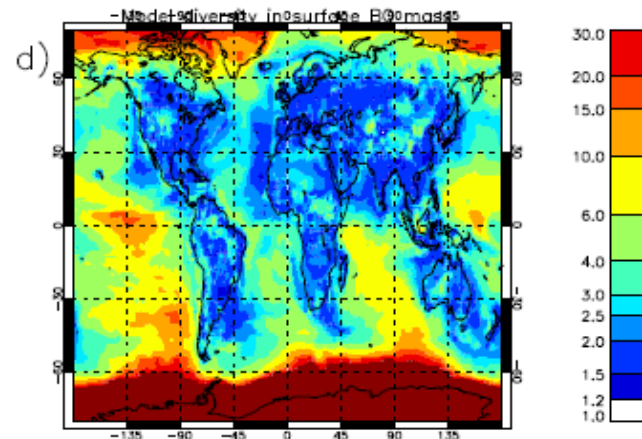
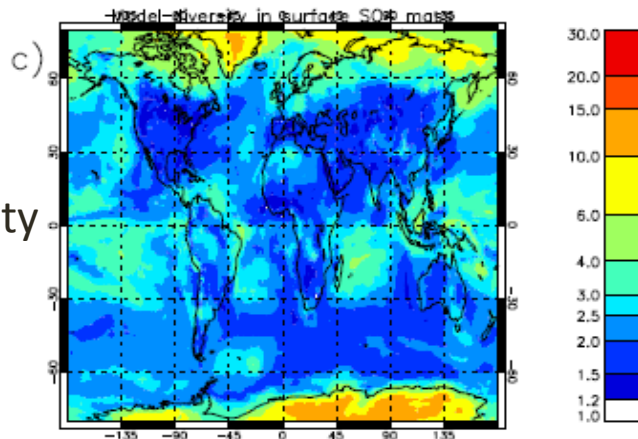
Central-8 model geometric mean

Model average



$E[Y]$

Model diversity

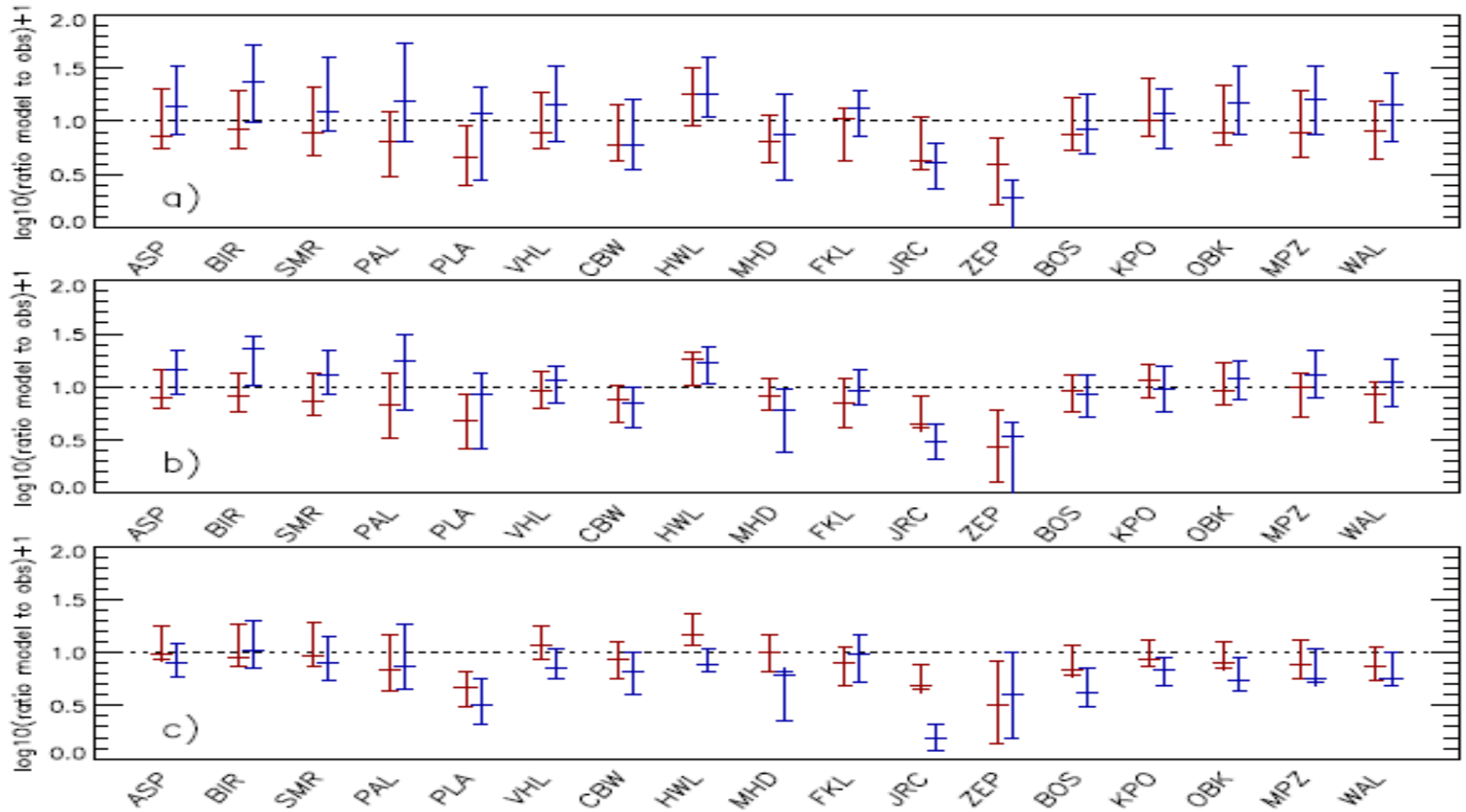


$[Y_{0.25}, Y_{0.75}]$

SO_4 concentration

Black carbon mass

AEROCOM II



Problems with MMEs

Ensemble of opportunity

Because the experiment is not *designed* to target particular sources of uncertainty it is difficult to know what is causing model diversity

x is limited and G_X is not fully explored, $P(Y|X)$ is not defined

Models are not independent

Models share code as people move modelling groups or collaborate. Some modelling centres also have multiple models

Have less degrees of freedom than assumed by independence

Single model runs

Modelling groups only submit their 'best' model run

Individual model uncertainty remains unexplored

Common errors

All models tend to represent the same processes depending on current scientific understanding
Comparison to observations will contain similar biases

Discussion: communicating structural uncertainty

Probability

Do people really think probabilistically despite the limited ensembles? How do they use them effectively? Similar issue with scenarios

Presenting ensembles

When presented with a central value and spread is the natural way of thinking like the Gaussian distribution?

Useful results from MMEs

- Although it is difficult to define \mathbf{X} for different structures using different models can provide important information on model diversity
- Different metrics are used to represent \mathbf{Y} helping to understand what is well/poorly modelled given current understanding
- Regional/temporal comparison of MME can help to identify processes with large structural diversity therefore areas for model development or the need for further understanding

Ways forward for MMEs

- Related models are removed from the ensemble
 - Dependence between models is reduced
- Discrepancy can be used in the statistical modelling to account for common biases
- Combine PPE and MME studies to account for and compare multiple sources of uncertainty

Discussion: two paradigms for MMEs

Truth + error

The model runs all represent reality with some error

As more models are added the uncertainty reduces

None of the models are reality

Exchangeability

Reality could be any one of the model runs

Model uncertainty is some representation of model spread

The models can simulate reality

In any case.....but justifiable?

Multi-model ensembles agree better with observations when the average is used.

Weighted averaged can be better if the weights can be calculated.

Perturbed physics ensembles - PPES

- Target parameter uncertainty
- Parameters are values used within the code to represent something in reality – sometimes measurable
- Parameter values chosen to give model output closest to observations
- Additional parameters added often without considering whether the definition of other parameters may have been changed by the additional model development.
- At any point in time a parameter is probably doing more than it was designed to do - we know that there are things happening in the real system that can't be modelled with current technology.

X and Y in a GLOMAP PPE

- In general, **Y** denotes all model outputs
 - From now on we will only consider a **single scalar model output, Y**
 - We consider only monthly mean cloud condensation nuclei (CCN) in a single model grid box
 - This simplifies the statistics
- In general, **X** denotes all model inputs, drivers, parameter values, initial conditions
 - Unless specified otherwise we use **X to denote model parameters and scaling of some inputs** – hereon in called model inputs
 - $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ with n inputs
 - X_i is a specific uncertain model input

$$Y = f(\mathbf{X})$$

We have added parameters to perturb model inputs that are not traditionally considered parameters

Aerosol model notation

Small letters are used to denote realisations of X and Y

$$\mathbf{x} = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$$

y = Monthly mean CCN concentration in each grid box

```
# Perturbation Settings (Initial Experiment)
# -----
```

```
ACT=20.5e-9
SE_SOL=0.0285
SE_INS=0.0285
NCRITFAC=0.259
NMOL=5.04e1
ANTHPARFRAC=0.000102
EMFRAC=0.704
EMSSFRAC=0.177
NSCAVACT=41.0e-9
```

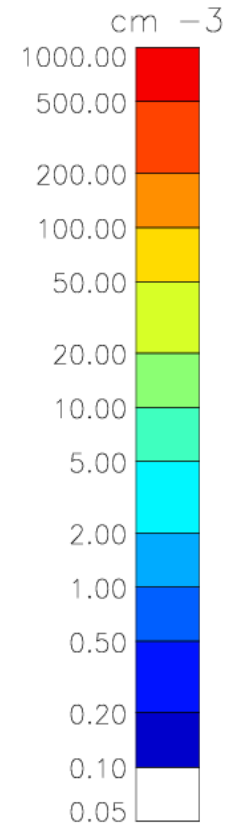
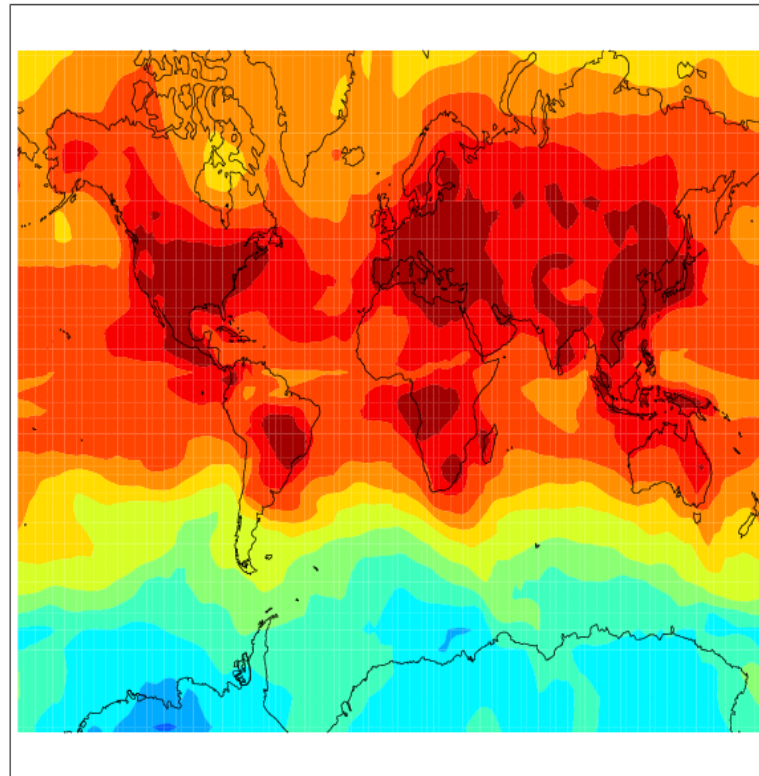
f

```
# Run The Model

echo "Running The Model"

if (test $MACHINE = "arc1")
then

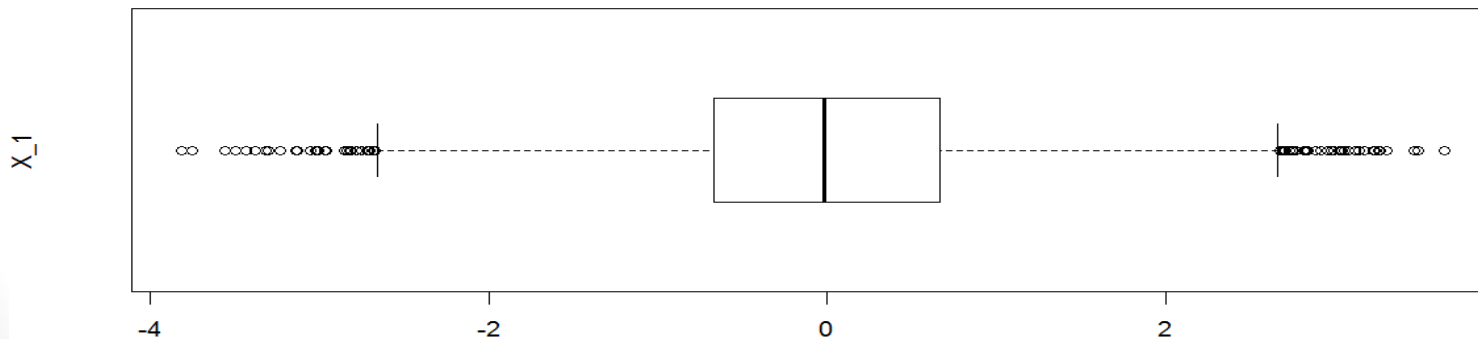
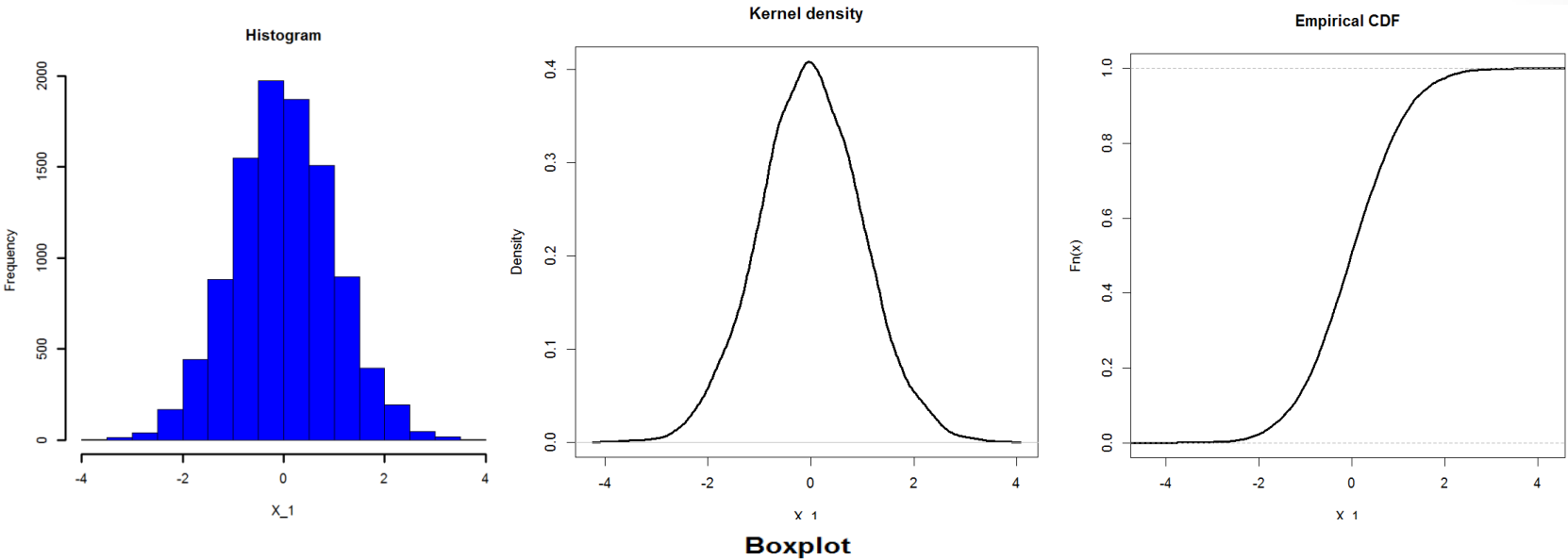
mpirun ./glomap.exe
```



Expert elicitation – what is the parameter uncertainty G_{X_i} ?

- A subjective method but in uncertainty analysis actually aimed at removing subjectivity
- Experts are encouraged to discuss the model and its parameters
- Encourage experts to push the ranges on the parameter values that they would normally choose
- Document all evidence and expert opinions that lead to G_{X_i}
- A statistician is present to avoid common problems such as anchoring and to form the probability distributions

Finding the probability distributions for X



```
> summary(x)
```

```
   Min.   1st Qu.   Median     Mean   3rd Qu.     Max.     
-3.814000 -0.670600 -0.013300 -0.003629  0.665500  3.648000
```

Quantiles to elicit the probability distributions

- The quantiles may be used to describe data or uncertainty
- The 100 quantiles are called the **percentiles**

- **Median** - 50th percentile
- **Interquartile Range (IQR)** – 25th – 75th percentile

- Can be used alone for description
- If a probability distribution is required you can try to match the percentiles
- Percentiles seem to be more intuitive and easier to explain

SHELF - The Sheffield Elicitation Framework

The Sheffield Elicitation Framework

SHELF v2.0

ELICITATION RECORD – Part 2 – Distribution

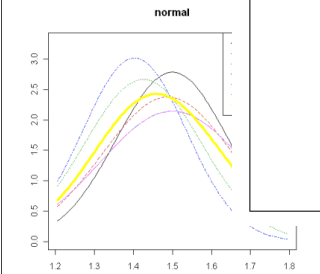
Quartile Method

Elicitation title	X4
Session	Experts - KC, JP, GM, KP, DS, PS
Date	01.07
Quantity	AIT_WIDTH
Start time	11.18

Definition	Modal Width – Accumulation mode width sigma						
Evidence	Philip uses 1.59 Heizenberg in Doc, 1.3 to 1.9 with mean 1.4						
Plausible range	1.2 to 1.8						
Median							
Upper and lower quartiles		KC	JP	GM	KP	DS	PS
	L	1.2	1.2	1.2	1.2	1.2	1.2
	Q1	1.4	1.35	1.33	1.3	1.35	1.35
	M	1.5	1.5	1.40	1.4	1.5	1.5
	Q3	1.6	1.6	1.55	1.5	1.65	1.65
	U	1.8	1.8	1.8	1.8	1.8	1.8

The Sheffield Elicitation Framework

Fitting



quantiles best.fit

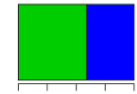
1	0.25	1.342351
2	0.33	1.380986
3	0.50	1.462800
4	0.66	1.544153
5	0.75	1.594479

The Sheffield Elicitation Framework

SHELF v2.0

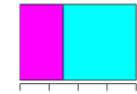
Group elicitation

lower quartile: 1.4



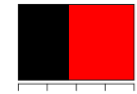
1.2 1.3 1.3 1.4 1.5

upper quartile: 1.6



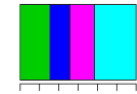
1.5 1.5 1.6 1.7 1.8

median: 1.5



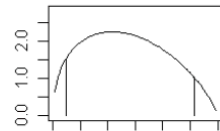
1.2 1.4 1.5 1.6 1.8

Four equally probable intervals



1.2 1.4 1.6 1.8

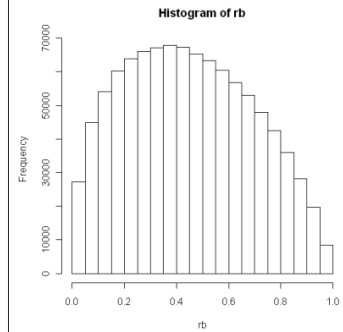
Sum of squares: 0.000118
0.05 quantile: 1.2
0.95 quantile: 1.7



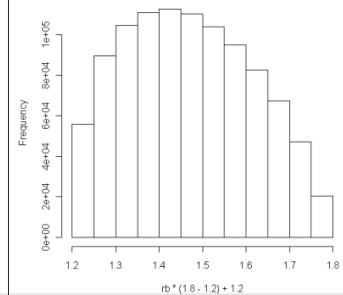
1.2 1.4 1.6 1.8
Scaled Beta(1.44 , 1.77)
mean = 1.47 , sd = 0.146

The Sheffield Elicitation Framework

Fitting and feedback



Histogram of $rb \cdot (1.8 - 1.2) + 1.2$



SHELF v2.0

	Min. 1st Qu. Median Mean 3rd Qu. Max.
	1.200 1.350 1.462 1.469 1.582 1.800
Chosen distribution	Beta(1.44,1.77) on the standardised scale.
Discussion	

Marginal and joint probability

$$Y = f(\mathbf{X})$$

- Each model input X_i has a pdf G_{X_i} - the marginal distribution of X_i
- Any combination of X_i and X_j has a joint probability distribution $G_{X_{i,j}}$
- **Very difficult to consider the joint uncertainty space a priori** but experts are encouraged to think about interactions between parameters
- We generally want to know the marginal G_Y given the G_X
- Interest also in G_Y given G_{X_i} and G_Y given $G_{X_{i,j}}$ etc. to understand how uncertainty in model inputs leads to uncertainty in model output
- If things are Gaussian all the joint probabilities can be derived, otherwise we use sampling
- **In GLOMAP X_i are mostly not Gaussian – need sampling**

GLOMAP Elicitation

Parameter	Lower	Upper
BCOC mass emission rate (fossil fuel)	0.5	2.0
BCOC mass emission rate (biomass burning)	0.25	4.0
BCOC mass emission rate (biofuel)	0.25	4.0
Sea spray mass flux (coarse/acc)	0.2x	5.0x
SO2 emission flux (anthropogenic)	0.6x	1.5x
SO2 emission flux (volcanic)	0.5x	2.0x
Biogenic monoterpene production of SOA	5 Tg/a	360Tg/a
Anthropogenic VOC production of SOA	3Tg/a	160Tg/a
DMS mass flux	0.5x	3.0x
BCOC mode diameter (fossil fuel)	30 nm	80 nm
BCOC mode diameter (biomass burning)	50 nm	200 nm
BCOC mode diameter (biofuel)	50 nm	200 nm
Subgrid conversion of SO2 to SO4 ("primary SO4")	0%	1%
Mode diameter of "primary SO4"	20 nm	100 nm

Particle and precursor gas emission rates

Properties of emitted particles

Have decided on the uncertain inputs
 $X = \{X_1, X_2, X_3, \dots, X_{27}, X_{28}\}$

Use the marginals G_{X_i} to sample
 $G_X = \{G_{X_1}, G_{X_2}, \dots, G_{X_{27}}, G_{X_{28}}\}$
 assuming independence

Parameter	Lower	Upper
BL nucleation rate k[H2SO4]	4E-7	2E-04
FT nucleation rate (BHN)	x0.01	X10
Ageing "rate" from insol to sol (monolayer)	0.3	5
Modal width (accumulation)	1.2	1.8
Modal width (Aitken)	1.2	1.8
Mode separation diameter (nucleation/Aitken)	9 nm	20 nm
Mode separation diameter (Aitken/accumulation)	x1.5	x3

Microphysical rates

Model "structural choices"

Cloud drop activation dry diameter	30	100
Reaction SO2 + O3 in cloud water (clean)	pH=4	pH=6.5
Reaction SO2 + O3 in cloud water (polluted)	pH=3.5	pH=5

Cloud processing

Nucleation scavenging dry D (above activation)	0 nm	100 nm
Nucleation scavenging fraction (T> -15C)	0.2	0.99
Dry deposition velocity (Aitken)	x0.5	X2.0
Dry deposition velocity (accumulation)	X0.1	X10.0

Dry and wet deposition

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, 8879-8914, doi:10.5194/acp-13-8879-2013, 2013.

Discussion points: elicitation

How much is expert judgement used, without explicit declaration?

What is the alternative to expert elicitation?

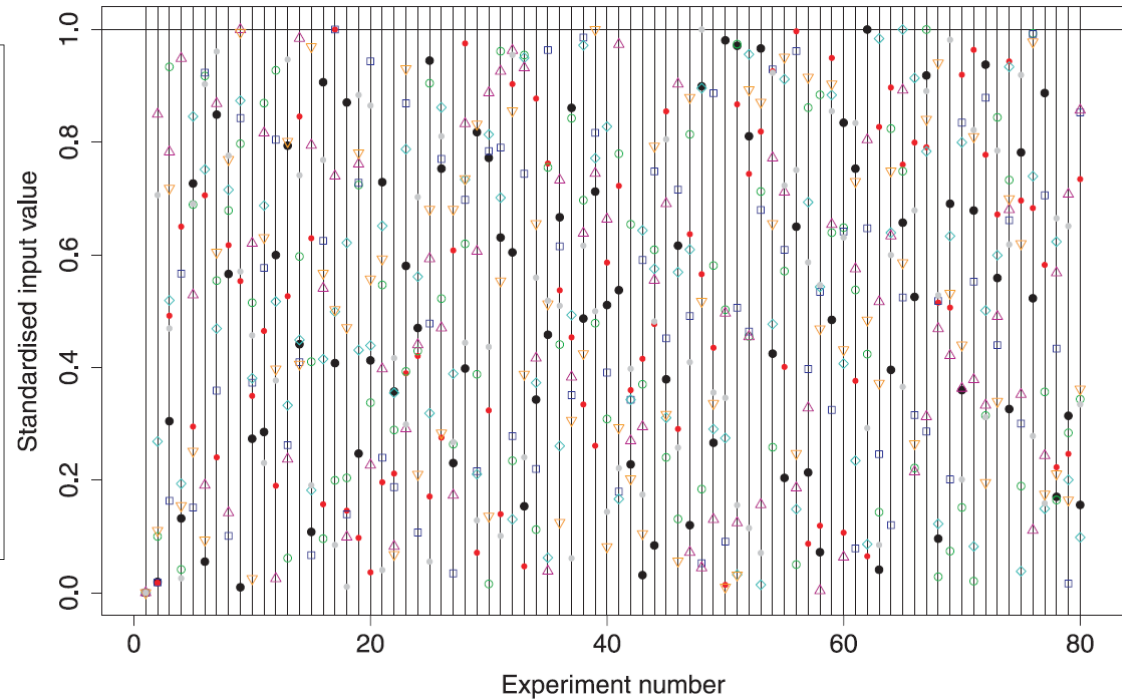
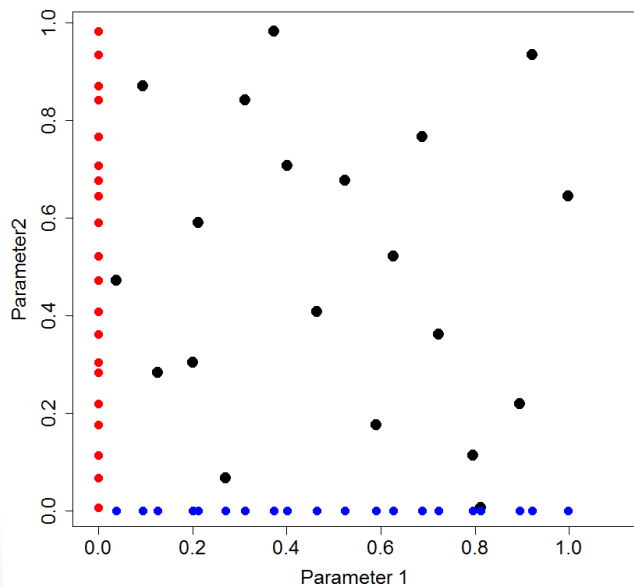
Expert elicitation is subjective, how subjective are objective statistics?

Experimental design for uncertainty

- In GLOMAP can't brute force sample from all probability distributions
- Aim to find maximum information in the fewest runs
- Different designs include:
 - Random sampling – poor space-filling properties
 - Factorial design – quickly becomes very large
 - Maximin latin hypercube – good space-filling and marginal properties
 - Sobol sequence – good space-filling properties
 - Hybrid designs – mixture of designs to target regions
 - 'Targetted' design – aim to reduce particular variances in an ensemble
- With GLOMAP we use the maximin latin hypercube implemented via R.

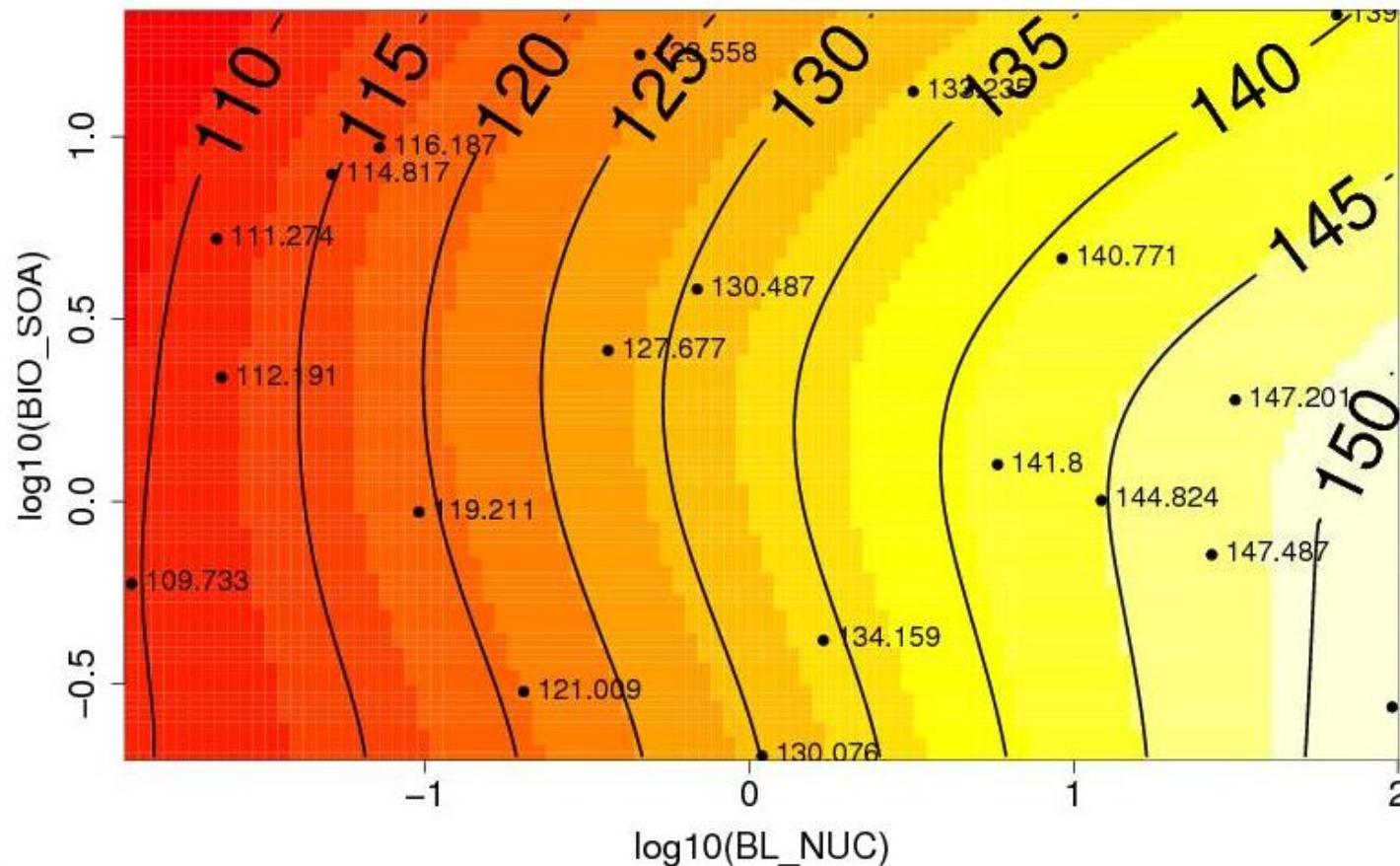
Experimental design II

- First step – space fill with latin hypercube
- If necessary target regions of the uncertainty space

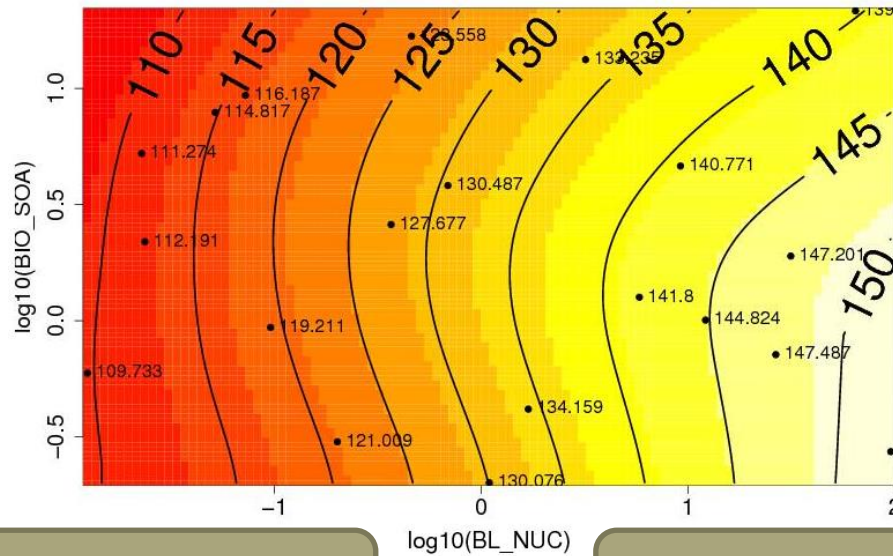


Emulation

Fill in the gaps between the limited runs
– large sample necessary

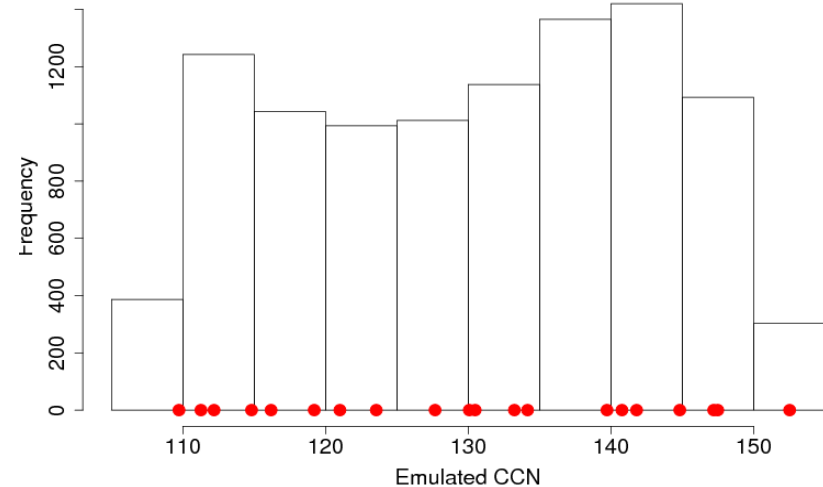
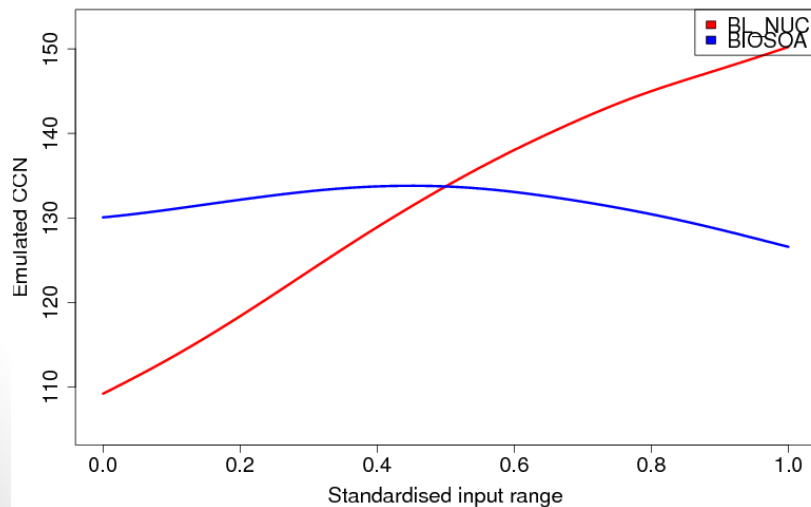


Emulation II

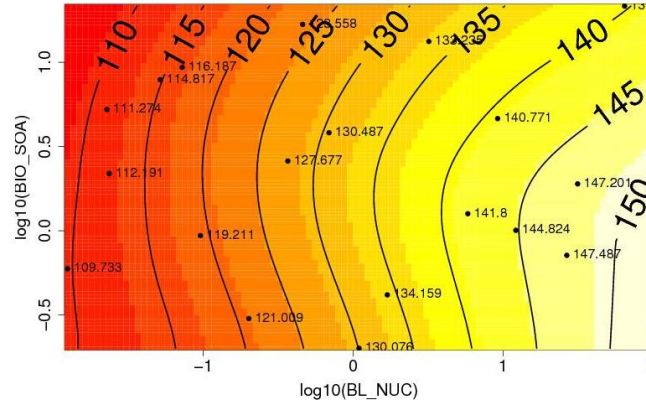


Find marginal relationships

Build up output distributions

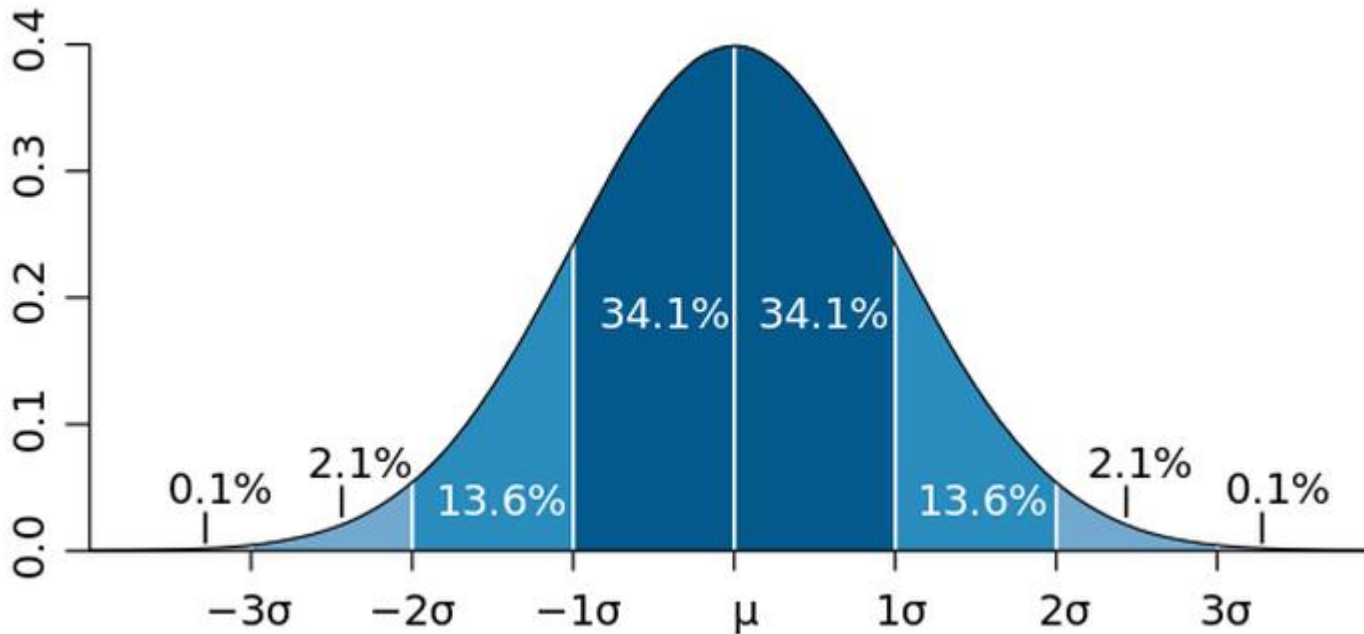


Emulation III – the Gaussian process



- Non-parametric but takes advantage of the conditional multivariate mathematics and Bayes Theorem
- Each model run is assumed to have a Gaussian distribution with zero variance - **the marginal**
- Together all model runs have a multivariate Gaussian distribution whose mean, variance and covariances can be calculated – **the joint**
- All unknown model runs have a conditional multivariate Gaussian distribution which can be calculated – **the conditional**
- We sample from this conditional distribution and use these to estimate GLOMAP where we don't have runs
- The variance in the conditional distribution gives us a measure of emulator uncertainty

Why is Gaussian so popular?



- The Gaussian (normal) is convenient mathematically
- It is sensible in a lot of cases
- Other distributions can be transformed to Gaussian
- The mean and median are the same, or at least similar
- Conditional and multivariate Gaussians are also Gaussian so still mathematically convenient

Emulation IV – the maths

The Gaussian process prior – specified by the mean and covariance functions

- The mean function – the output is some function of the inputs

$$E\{f(\mathbf{x})|\boldsymbol{\beta}\} = \mathbf{h}(\mathbf{x})^T \boldsymbol{\beta}, \quad \mathbf{h}(\cdot) = (\mathbf{1}, \mathbf{x}^T)$$

- The covariance function – the output is some function of the inputs with error dependent on the distance between points and the output ‘smoothness’

$$\text{cov}\{f(\mathbf{x}), f(\mathbf{x}')|\sigma^2\} = \sigma^2 c(\mathbf{x}, \mathbf{x}') \\ c(\mathbf{x}, \mathbf{x}') = \exp\{-(\mathbf{x} - \mathbf{x}')^T \mathbf{R}(\mathbf{x} - \mathbf{x}')\}$$

- The parameters in the statistical function will be estimated by the known points (training data)

$$p(\boldsymbol{\beta}, \sigma^2) \propto \sigma^2$$

Emulation V – the maths

The Gaussian process posterior – also specified by the mean and covariance functions

- The mean function goes through the training points

$$m^*(\mathbf{x}) = \mathbf{h}(\mathbf{x})^T \hat{\boldsymbol{\beta}} + \mathbf{t}(\mathbf{x})^T \mathbf{A}^{-1} (\mathbf{y} - \mathbf{H} \hat{\boldsymbol{\beta}})$$

- The covariance function – there is zero error at the training points and it gets bigger as you move away from them

$$\hat{\sigma}^2 c(\mathbf{x}, \mathbf{x}')^* = \hat{\sigma}^2 (c(\mathbf{x}, \mathbf{x}') - \mathbf{t}(\mathbf{x})^T \mathbf{A}^{-1} \mathbf{t}(\mathbf{x}') + (\mathbf{h}(\mathbf{x})^T - \mathbf{t}(\mathbf{x})^T \mathbf{A}^{-1} \mathbf{H}) (\mathbf{H}^T \mathbf{A}^{-1} \mathbf{H})^{-1} (\mathbf{h}(\mathbf{x}')^T - \mathbf{t}(\mathbf{x}')^T \mathbf{A}^{-1} \mathbf{H})^T)$$

Emulation VI – the maths

The Gaussian process posterior – the dependency on the training data

$$\mathbf{H}^T = (\mathbf{h}(\mathbf{x}_1), \dots, \mathbf{h}(\mathbf{x}_n)),$$

$$\mathbf{A} = \begin{pmatrix} 1 & c(\mathbf{x}_1, \mathbf{x}_2) & \dots & c(\mathbf{x}_1, \mathbf{x}_n) \\ c(\mathbf{x}_2, \mathbf{x}_1) & 1 & & \vdots \\ \vdots & & \ddots & \\ c(\mathbf{x}_n, \mathbf{x}_1) & \dots & & 1 \end{pmatrix},$$

Used
DiceKriging
in R

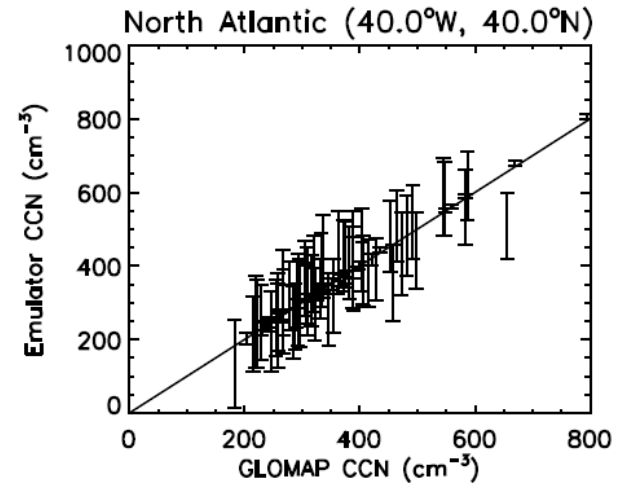
$$\hat{\boldsymbol{\beta}} = (\mathbf{H}^T \mathbf{A}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{A}^{-1} \mathbf{y}$$

and

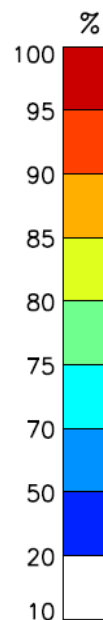
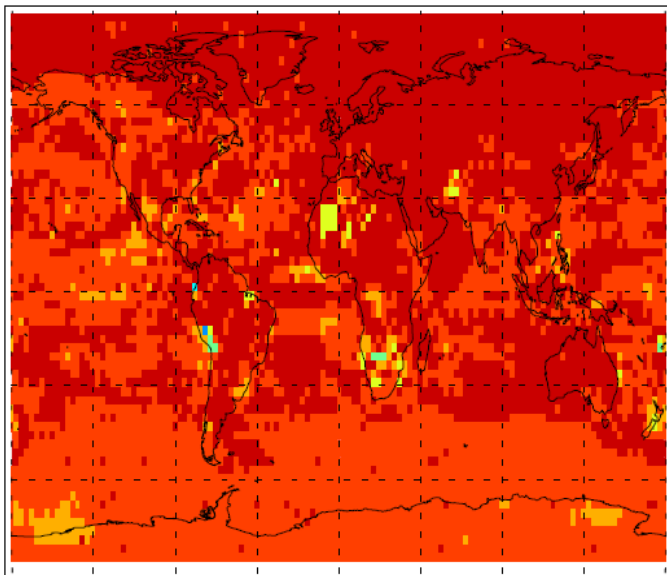
$$\hat{\sigma}^2 = \frac{\mathbf{y}^T (\mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{H} (\mathbf{H}^T \mathbf{A}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{A}^{-1}) \mathbf{y}}{n - q - 2}$$

Emulator validation

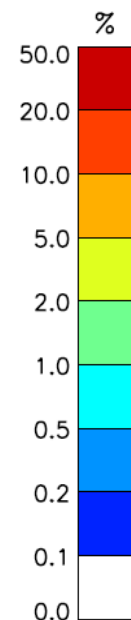
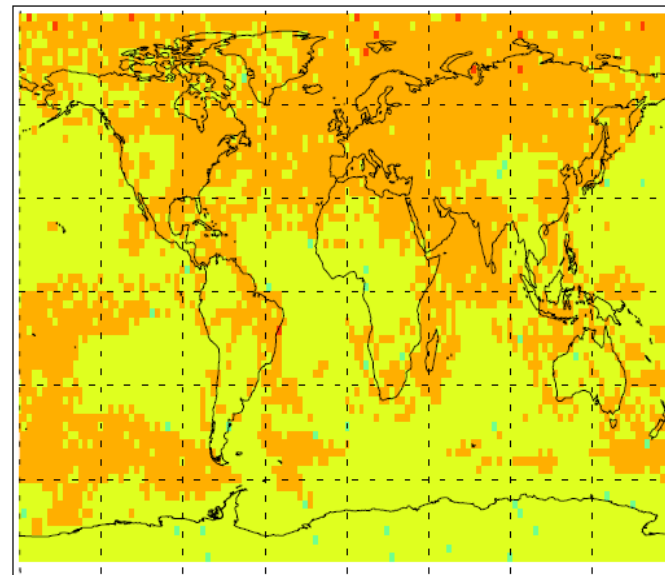
Are the emulator estimates close to the GLOMAP points?



a) JAN emulator validation



b) JAN emulator uncertainty



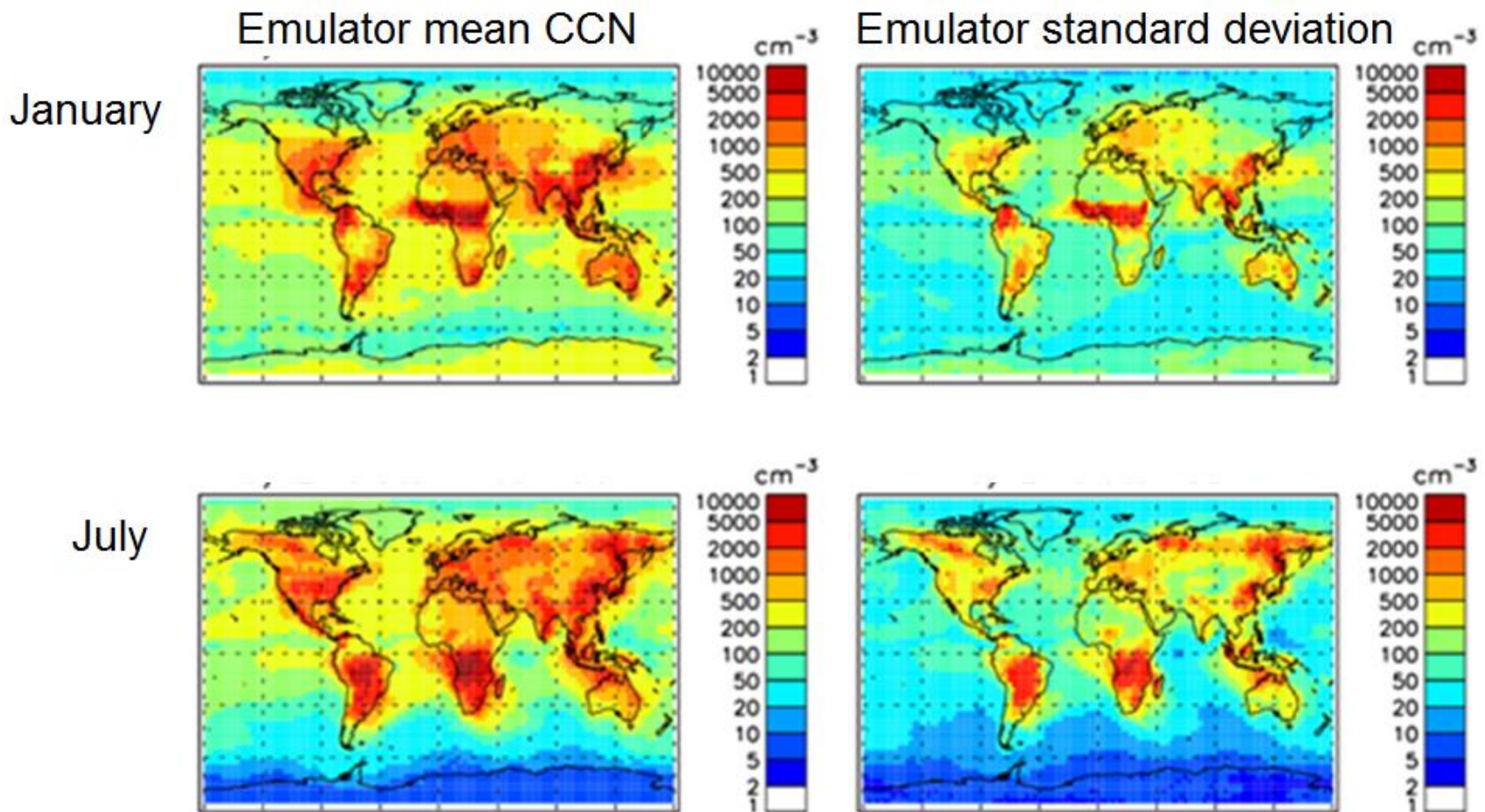
Uncertainty analysis (UA)

- What is the uncertainty in the output due to the uncertain inputs?
- What is the variance of $E[Y|X]$?
- We sample from G_X over all X and use the emulator to calculate the associated variance in Y

Sample 140,000 emulator runs for
GLOMAP UA and SA – would take
~190 years with GLOMAP

UA II

- Carried out UA on each of the model grid boxes separately and plotted in a map



Sensitivity analysis

- Uncertainty in which parameter leads to most uncertainty in the model output? Which G_{X_i} variance results in the biggest portions of G_Y variance?

$$\text{Var}(Y) = \sum_i W_i + \sum_{i < j} W_{i,j} + \dots + \sum_i W_{1,2,\dots,p}$$

$$V_i = W_i, \quad V_{i,j} = V_i + V_j + W_{i,j}$$

$$S_i = \frac{V_i}{\text{Var}(Y)}$$

$$\sum_{i=1} S_i + \sum_{i < j} S_{ij} + \dots + S_{12\dots p} = 1$$

SA – main effect and total effect

- **Main effect** – percentage of variance that could be reduced if we learn X_i (or by how much it's irreducible if it can't be learnt)

$$S_i = V_i / \text{var}(Y)$$

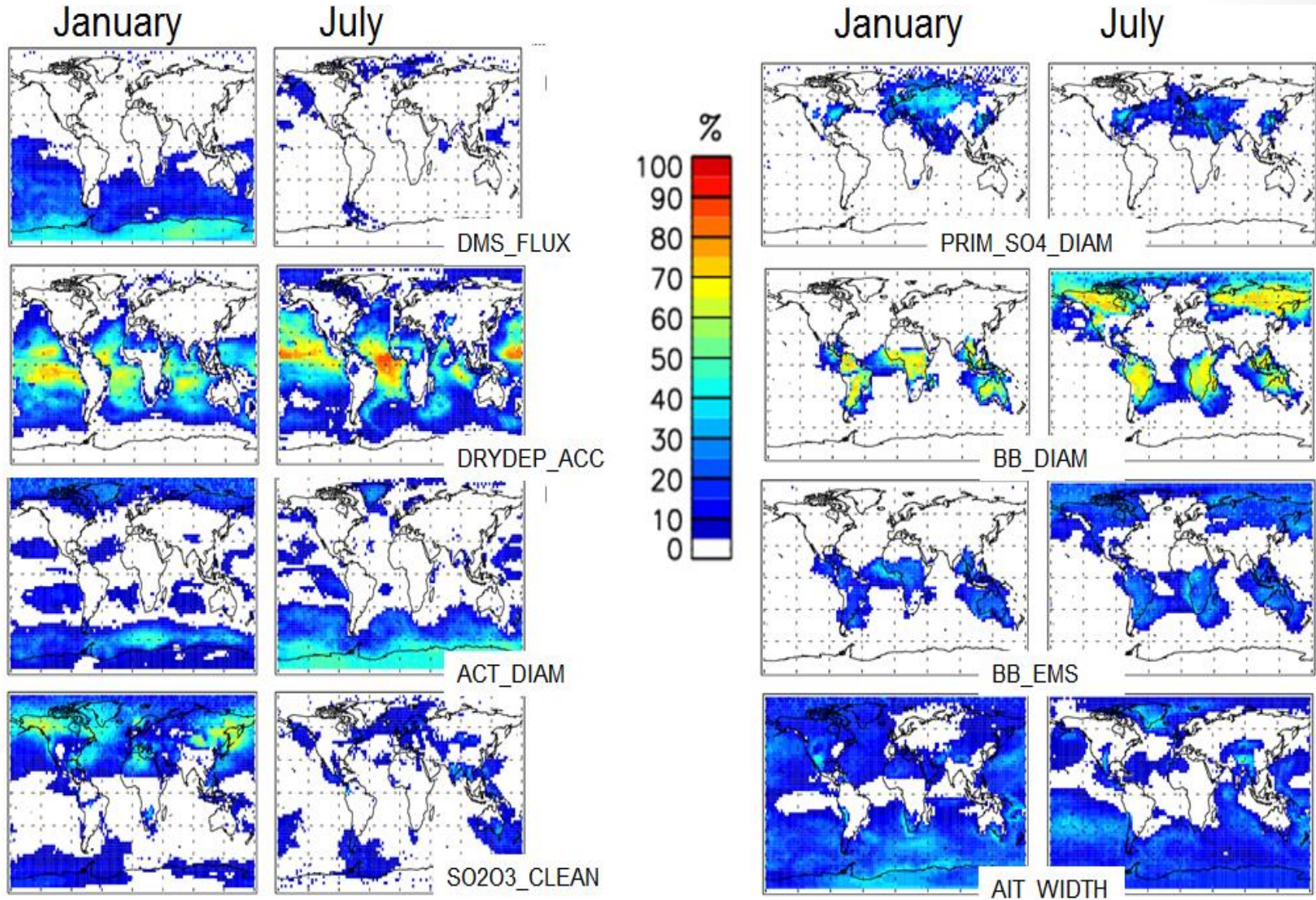
- **Total effect** – percentage of variance that remains unexplained if we learn everything but X_i

$$S_{Ti} = V_{Ti} / \text{var}(Y) = 1 - S_{-i}$$

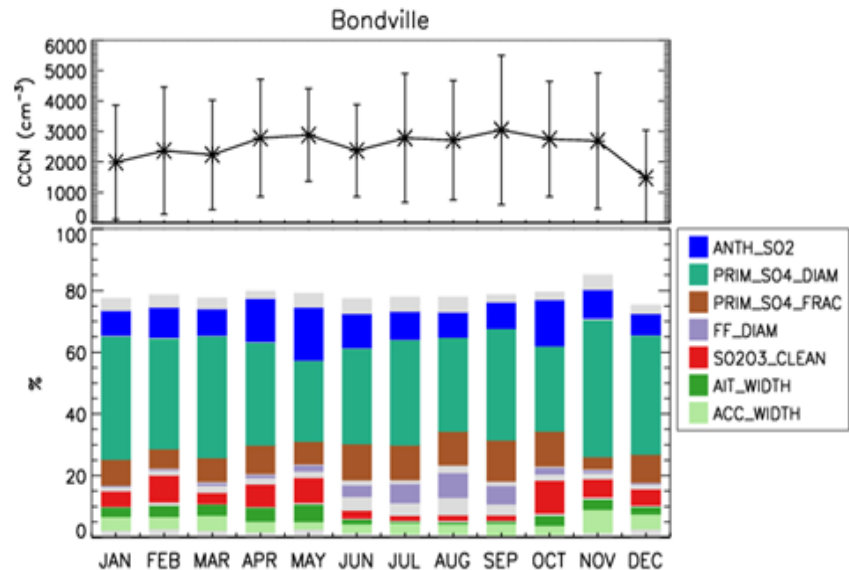
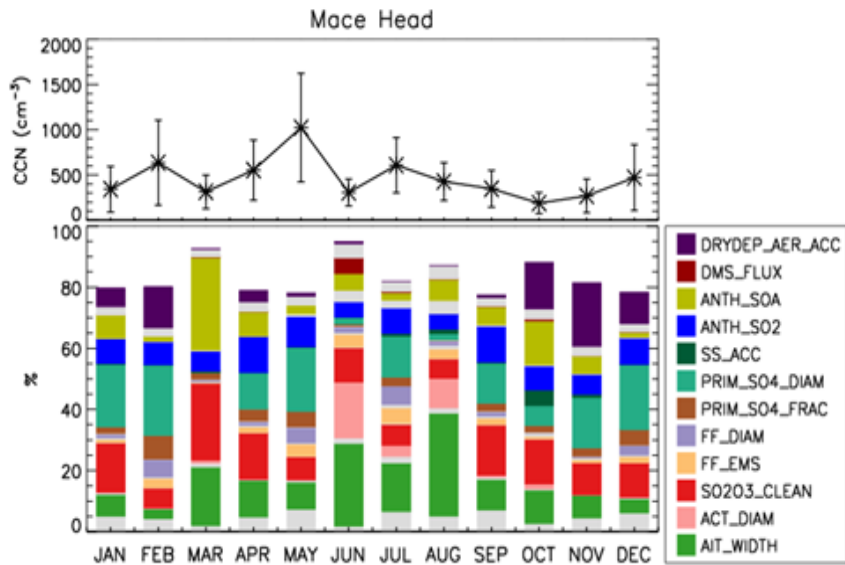
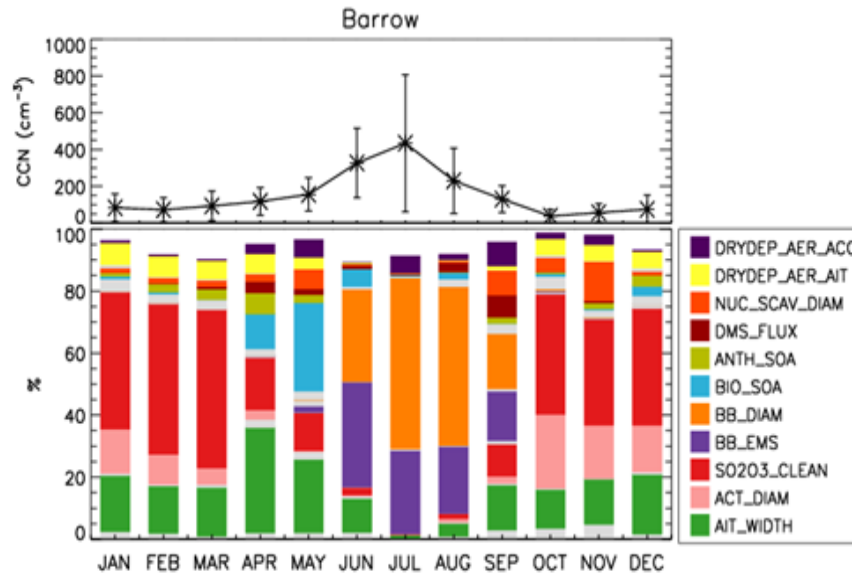
- By comparing main and total effect we learn about interactions

Used sensitivity in R

SA II – mapping the results



SA III – time series of results



Learning about computer models

- Visualising the results over space and time can help to learn about the computer model
- Even without observations we have learnt about the model
- How does the model behaviour compare to what we know scientifically?
- Are the models of any use? Adequate for purpose
- Using models to predict – more justifiable if we know the model behaviour is consistent with science and responds as expected

Learning about GLOMAP

- The model responds to perturbations in multiple inputs as expected from current knowledge
- The model responses show emissions transport consistently with current knowledge
- The model seasonality is consistent with current knowledge
- We have also learnt which of the perturbations lead to the largest uncertainties and which uncertainties need to be improved
- We have also learnt which model perturbations don't cause a significant model response

Ways forward for PPE

- More model outputs to be studied
- Better emulators for multivariate data?
- Observationally constrained PPEs

Comparing models and observations

- Learnt about GLOMAP but how consistent is it with reality?
- Is the model at all consistent with observations?
- What metrics should be used?
- How many metrics help avoid compensating errors?
- How good is good enough?
- Can a model just be fit for purpose?

History matching

Implausibility measure for a single model output:

$$\mathcal{I}_i(x)^2 = \frac{(z_i - \mathbb{E}[f_i(x)])^2}{\text{Var}[z_i - \mathbb{E}[f_i(x)]]}$$

Choose a threshold and rule out any runs in which I exceeds the threshold


Implausibility measure for multivariate model output:

$$\mathcal{I}(x) = (z - \mathbb{E}[f(x)])^T \text{Var}[z - \mathbb{E}[f(x)]]^{-1} (z - \mathbb{E}[f(x)])$$

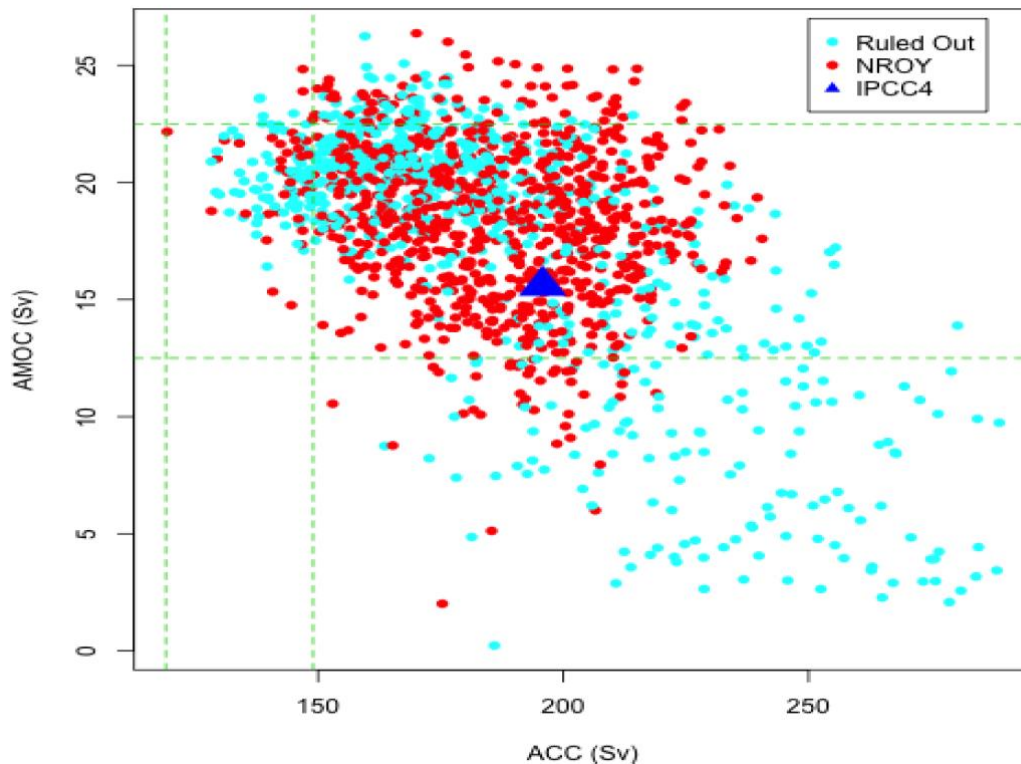
Include the emulator variance, discrepancy variance and observation error:

$$\text{Var}[z - \mathbb{E}[f(x)]] = \text{Var}[f(x)] + \text{Var}[\eta] + \text{Var}[e]$$

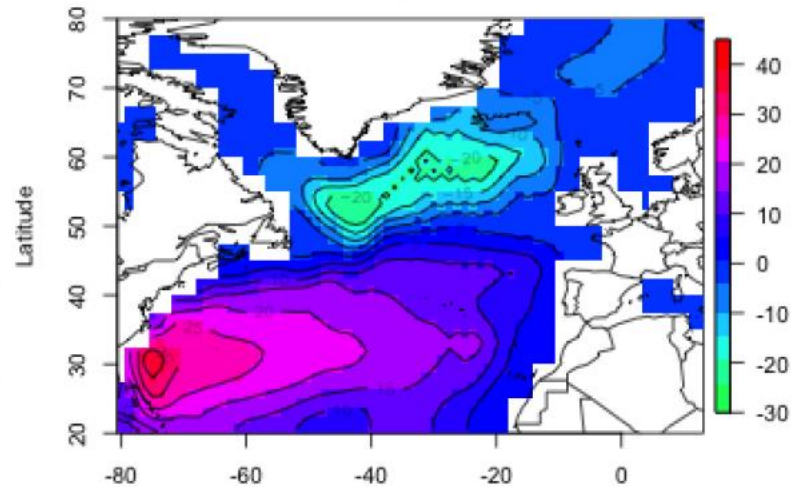
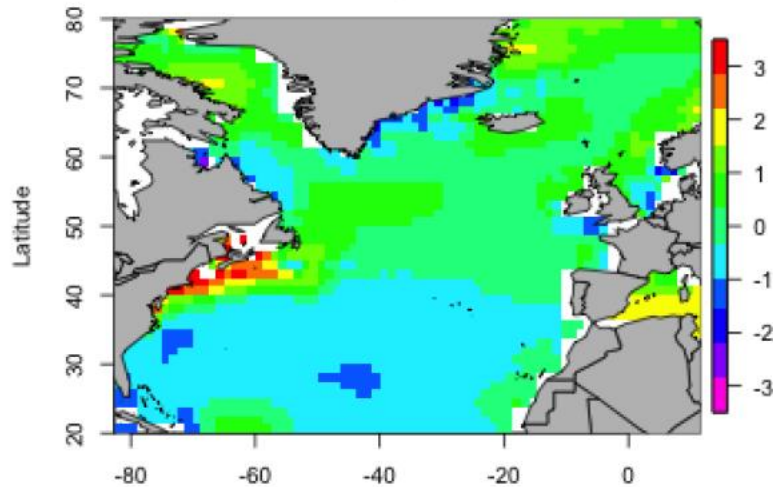
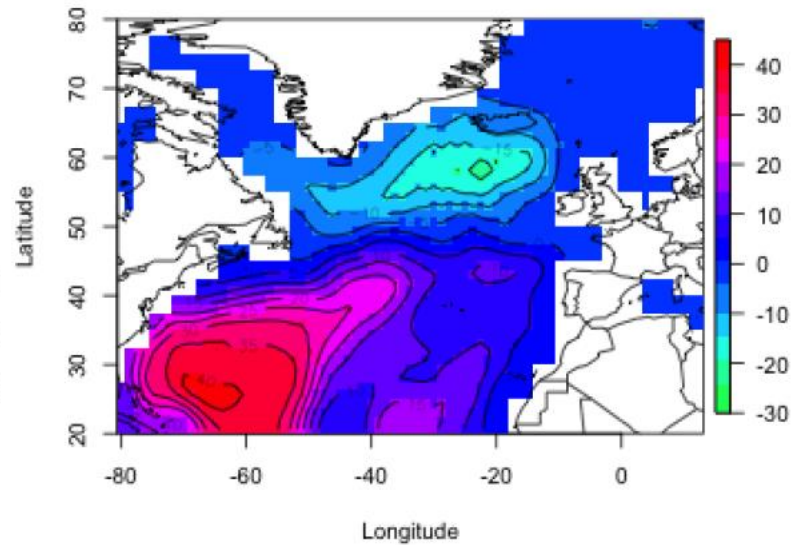
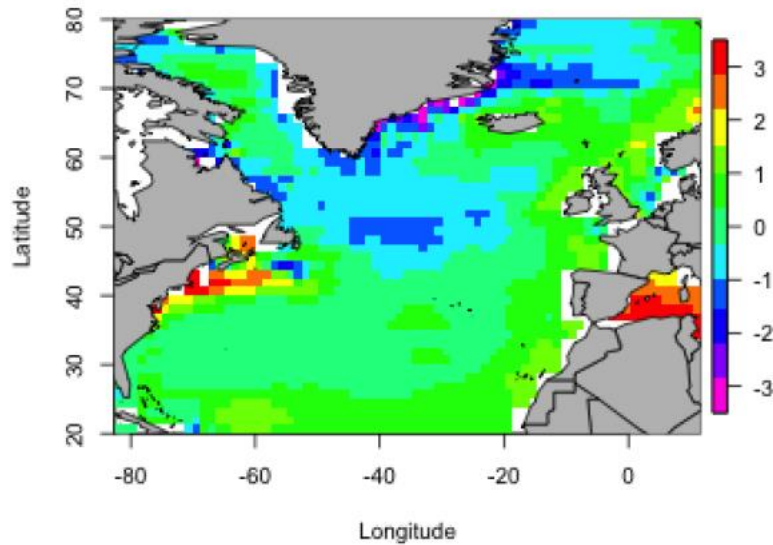
History matching HadCM3, using climateprediction.net

 Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P. Jackson, L., Yamazaki, K. (2013),

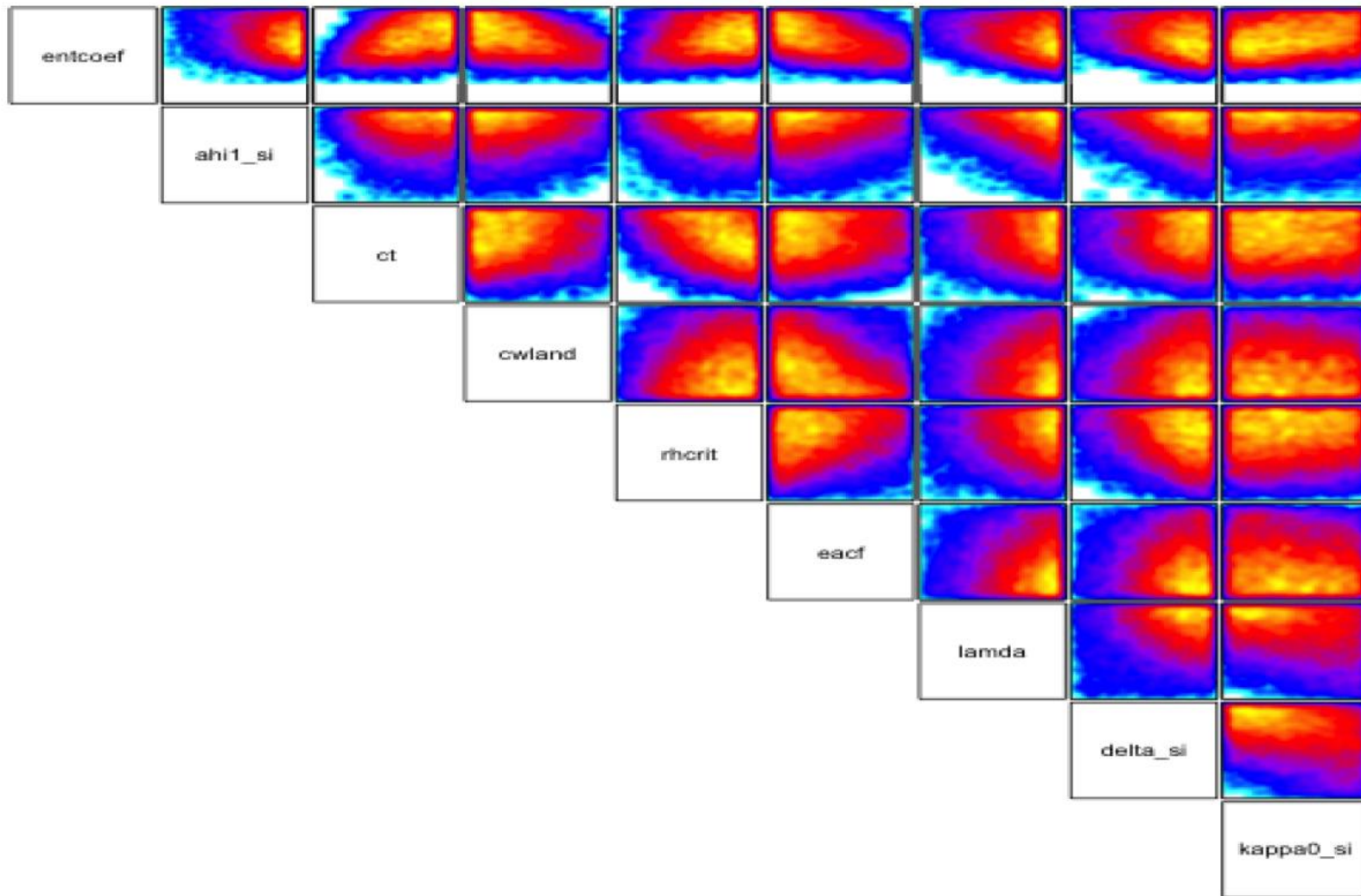
History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble, *Climate Dynamics*, 41(7), 1703–1729.



History matching



History matching



History matching

- We have found models with more realistic ocean circulations.
- History matching has lead us to a region of parameter space that might contain even better models.
- We further reduce this space using 5 further constraints:
 - 1 SST in the sub-tropical gyre **Space Reduction 82.5%**
 - 2 SSS in the sub-polar gyre **Space Reduction 29.8%**
 - 3 The STG is stronger west of 75°W than to the east at 30°N .
Space Reduction 0.4%
 - 4 The SPG is 1.5 times stronger in the labrador sea than to the west of Greenland. **Space Reduction 8.2%**
 - 5 SST around iceland. **Space Reduction 4.5%**
- The remaining space is 0.47% of the original space.



Calibration

- Find the ‘best’ model parameter values
- Must include discrepancy
- Discrepancy can be modelled as a Gaussian process
- Posterior introduced earlier becomes the prior and the observations are used to update it for a new observationally constrained posterior
- Can estimate the calibration parameters from the posterior taking into account parameter uncertainty, discrepancy and observational error

We represent the relationship between the observations \mathbf{z} , the true process $\zeta(\cdot)$ and the computer model output $\eta(\cdot, \cdot)$ in the equation

$$z_i = \zeta(\mathbf{x}_i) + e_i = \rho\eta(\mathbf{x}_i, \boldsymbol{\theta}) + \delta(\mathbf{x}_i) + e_i, \quad (7)$$

where e_i is the observation error for the i th observation, ρ is an unknown regression parameter and $\delta(\cdot)$ is a model inadequacy function that is *independent* of the code output $\eta(\cdot, \cdot)$.

Including observation error and discrepancy

- We know they exist so must account for them
- They are very hard to specify

Observation error

Include:

- measurement error
- scale differences (time and space)
- variability

How do we measure these?

Discrepancy

Models cannot simulate exact reality

How can we measure this and use the data to calibrate?

Discussion: what is the 'best' model?

History matching

Rule out parts of parameter space inconsistent with current set of observations
- is there a best model?

Calibration

Find parameters that match the current observations most consistently
- assumes there is a best model

Discussion: communication and education

How much statistics do you need to understand?

Do you need to understand all aspects of statistics to use them?

Who should be using them?

References

- **Craig, P. S., Goldstein, M., Seheult, A. H., and Smith, J. A. (1996), "Bayes Linear Strategies for Matching Hydrocarbon Reservoir History," in Bayesian Statistics 5, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., Oxford University Press, pp. 69-95.**
- Kennedy, M. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society, Series B.* 63, 425-464.
- **Knutti, R., R. Furrer, C. Tebaldi, J. Cermak and G.A. Meehl, 2010, Challenges in combining projections from multiple models, *Journal of Climate*, 23, 2739-2758, doi:10.1175/2009JCLI3361.1.**
- 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, 12253-12273, doi:10.5194/acp-11-12253-2011, 2011.
- **Lee, L. A., Carslaw, K. S., Pringle, K. J., and Mann, G. W.: Mapping the uncertainty in global CCN using emulation, *Atmos. Chem. Phys.*, 12, 9739-9751, doi:10.5194/acp-12-9739-2012, 2012.**
- 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, 8879-8914, doi:10.5194/acp-13-8879-2013, 2013.
- 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, 519–551, doi:10.5194/gmd-3-519-2010, 2010
- Oakley, J. and O'Hagan, A.: Probabilistic sensitivity analysis of complex models: a Bayesian approach, *J. Roy. Stat. Soc. B*, 66, 751–769, 2004.
- Oakley J. E. and O'Hagan, A. (2010). SHELF: the Sheffield Elicitation Framework (version 2.0), School of Mathematics and Statistics, University of Sheffield, UK. (<http://tonyohagan.co.uk/shelf>)
- O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. E., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E. and Rakow, T. (2006). *Uncertain Judgements: Eliciting Expert Probabilities*. Chichester: Wiley.
- **Saltelli, A., Chan, K., and Scott, M. E.: Sensitivity Analysis, New York, Wiley, 2000.**

