Introduction to emulators

- the what, the when, the why

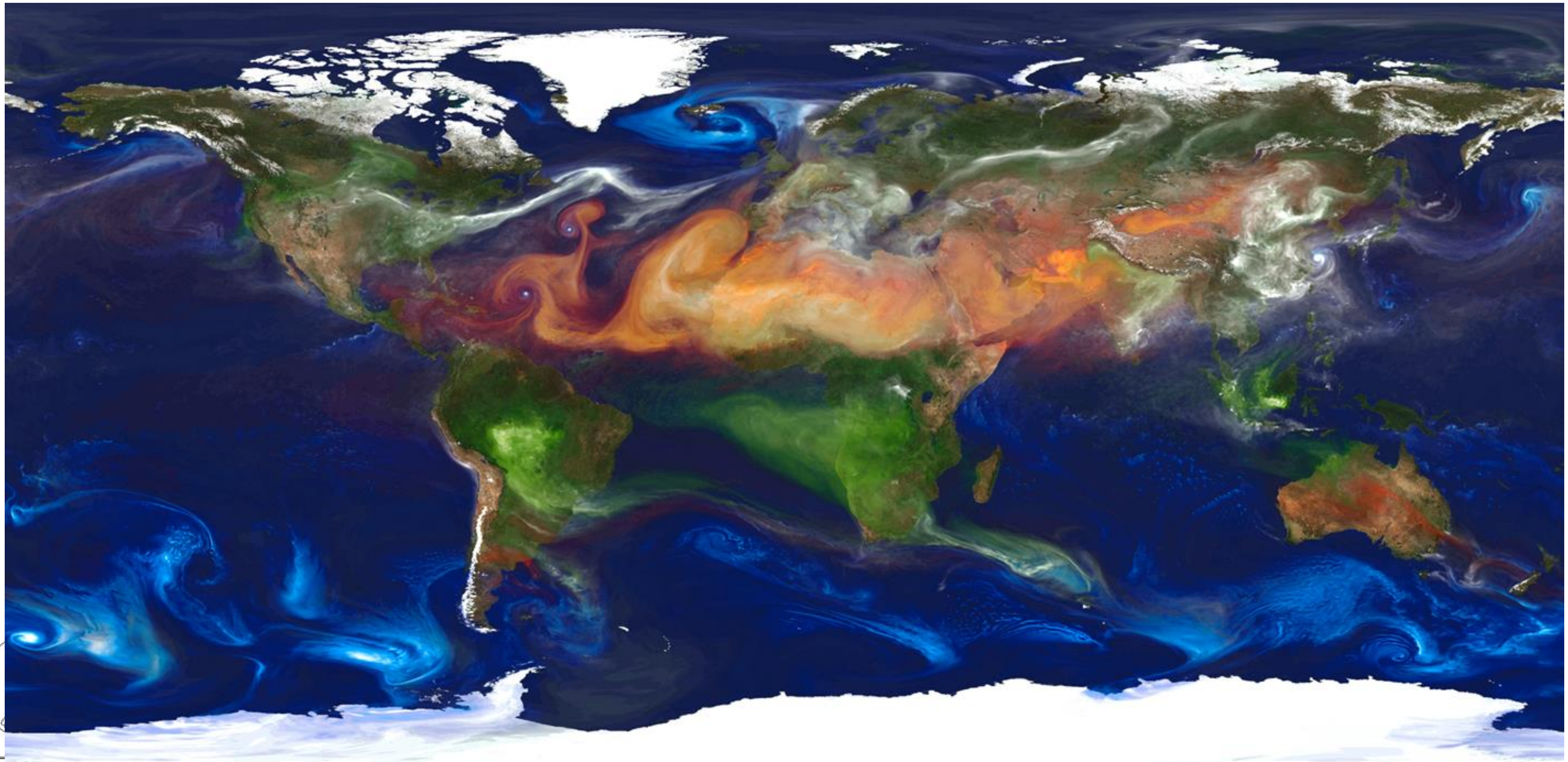
Dr Lindsay Lee

The Leverhulme Trust

What is a simulator?

UNIVERSITY OF LEEDS

A simulator is a computer code used to represent some real world process



The Leverhulme Trust

Why simulate?

UNIVERSITY OF LEEDS

Understand

Which aerosol are most effective at cooling the climate?

Predict

Could aerosol be used to cool the climate for geoengineering?

The Leverhulme Trust

Notation

- $y = f(x)$ represents the simulation process
- y is the simulator output
- x is the simulator input
- f is the simulator

Uncertainty in simulation

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

Computer codes are imperfect representations of the real world.

How do these imperfections affect our understanding and predictions?
What is the effect on $\boldsymbol{Y}$?

Uncertainty in climate simulators

UNIVERSITY OF LEEDS

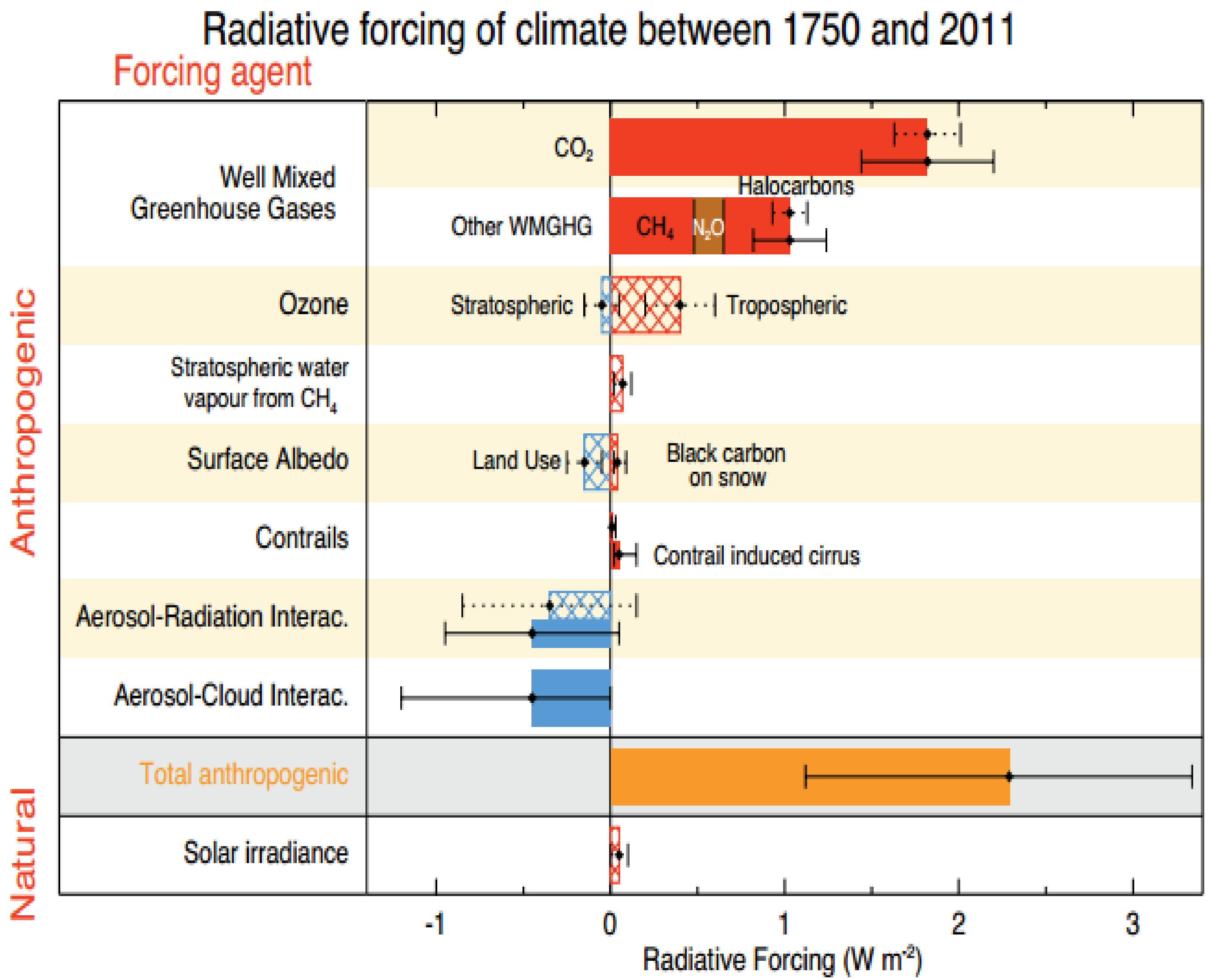


Figure 8.15 | Bar chart for RF (hatched) and ERF (solid) for the period 1750–2011, where the total ERF is derived from Figure 8.16. Uncertainties (5 to 95% confidence range) are given for RF (dotted lines) and ERF (solid lines).

The Leverhulme Trust

What is X?

- $\boldsymbol{X}$ are ‘model inputs/parameters’
- They may be spatial fields/time series
- They may be single values used globally or regionally

The Leverhulme Trust

Where X is a spatial field or series…

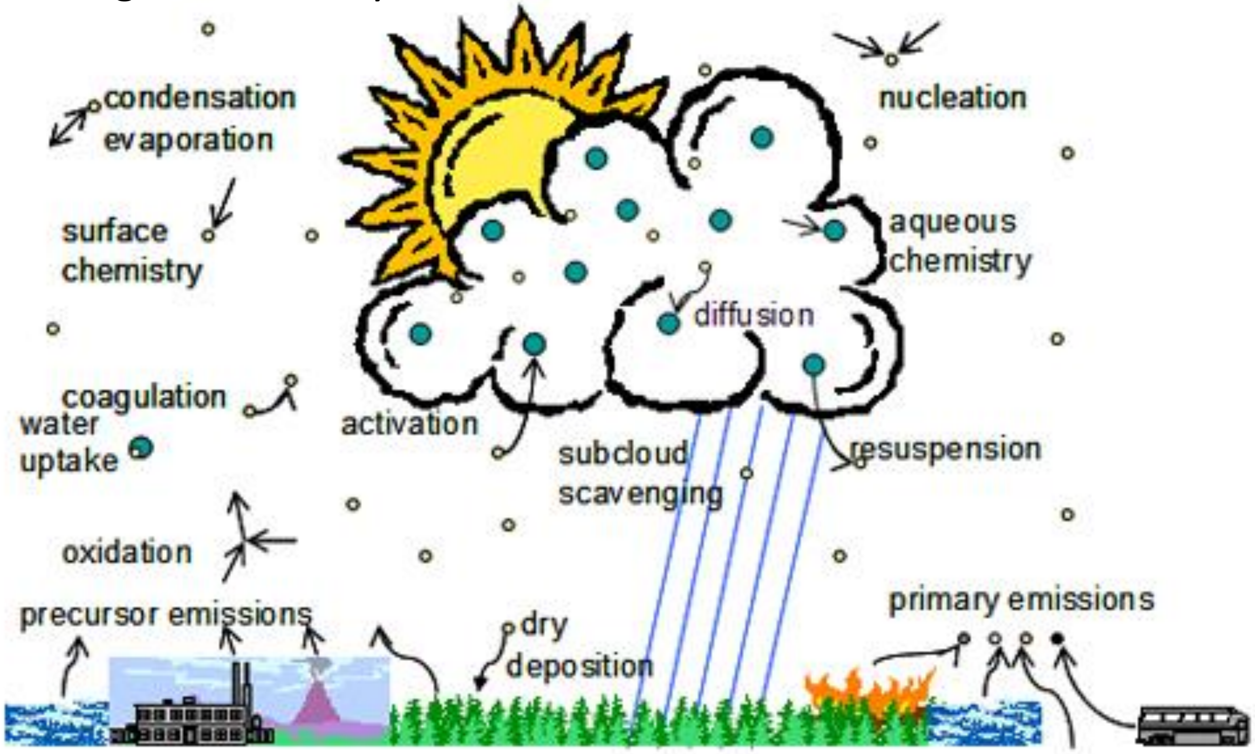
UNIVERSITY OF LEEDS

We will represent the uncertainty in $\boldsymbol{X}$ by a single value to perturb the whole field, or series (maybe regionally, but mostly globally)

The Leverhulme Trust

Uncertainty in aerosol simulator inputs

Diagram courtesy of PNNL



The Leverhulme Trust

Uncertainty in simulator inputs

UNIVERSITY OF LEEDS

- $\boldsymbol{X}$ is uncertain $\rightarrow$ $\boldsymbol{Y}$ is uncertain
- $\boldsymbol{x}$ can be given uncertainty distribution $\boldsymbol{G}$
- Marginally x_k has distribution G_k

Obtain $\boldsymbol{y}_i = f(\boldsymbol{x}_i)$ for given $\boldsymbol{x}_i$ from $\boldsymbol{G}$

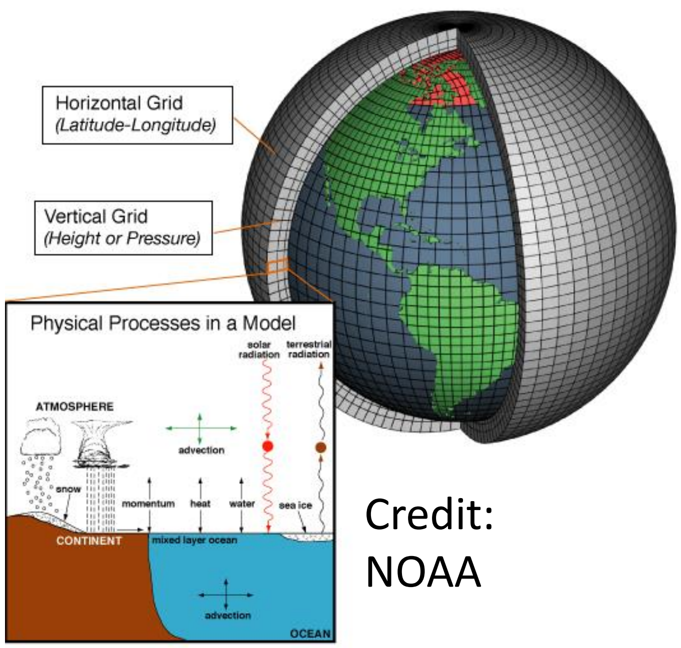
The Leverhulme Trust

Simulator complexity

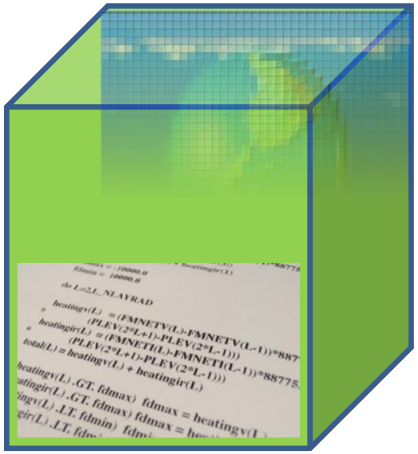
UNIVERSITY OF LEEDS

- Simulators contain lots of equations (hundreds of lines of code)

- Simulator resolution increases with computer power



Credit: NOAA



The Leverhulme Trust

Bottom line…

UNIVERSITY OF LEEDS

f is often too complex to simulate many $\boldsymbol{x}_i$ from $\boldsymbol{G}$ to characterise the effect of uncertainty on $\boldsymbol{y}$

The Leverhulme Trust

Emulation

UNIVERSITY OF LEEDS

- Replace f with $\hat{f}$
- $\hat{f}$ is the emulator
- $\hat{f}$ is much quicker to run but accurately represents f

The Leverhulme Trust

More accurately…

- Replace $y_k = f(x)$ with $y_k = \hat{f}(x)$

Define an output (or selection of) to be emulated

- doesn't replace the whole simulator for all research

What is $\mathbf{y}$?

- From the simulator $\boldsymbol{y}$ can be any/all model output/diagnostics
- For emulation $\mathbf{y}$ must be more restricted
- y is often a scalar representing a single model output, at a particular time and location
  - Eg.
    - global mean temperature for 2000
    - total number for particulate matter aerosol > 2.5um in a model grid box
- $\boldsymbol{y}$ could be a spatial field, or a time series, or EOF/PCAs, or a selection of model outputs

Emulator complexity

UNIVERSITY OF LEEDS

The complexity of the emulator is dependent on the complexity of the model output to be emulated

The Leverhulme Trust

Continuing…..

UNIVERSITY OF LEEDS

- We will carry on assuming y is scalar and $\boldsymbol{x}$ is a collection of scalar values

- $\boldsymbol{x} = x_1, x_2, x_3, \ldots, x_p$ for p uncertain parameters

- Use a selection of $\boldsymbol{x}_i, \boldsymbol{y}_i$ from f to 'build' $\hat{f}$

The Leverhulme Trust

When is an emulator useful?

UNIVERSITY OF LEEDS

uncertainty analysis

Quantify the impact of uncertainty on model output

sensitivity analysis

Understand model response to uncertain inputs

And the model is too complex to do the sampling required

The Leverhulme Trust

A statistical emulator

UNIVERSITY OF LEEDS

An emulator that gives estimates of its own uncertainty

We tend to use the Gaussian process

The Leverhulme Trust

The Gaussian process emulator,
(Rasmussen and Williams, 2006 & O'Hagan 2006).

UNIVERSITY OF LEEDS

- Each point in space has a Gaussian distribution
- Each collection of points has a multivariate Gaussian distribution
- The whole collection is a Gaussian process
- In its basic form requires 'smooth output'

Does not require the output to be Gaussian

The Leverhulme Trust

The GP prior

Set up the Gaussian process prior function

$$f(\boldsymbol{x})|\beta,\sigma \sim GP(h(\boldsymbol{x})^T\beta, \sigma^2 c(\boldsymbol{x},\boldsymbol{x}'))$$

$h(\boldsymbol{x})^T\beta$ is the mean function

$c(\boldsymbol{x},\boldsymbol{x}')$ is the covariance function

β,σ are hyperparameters

The Leverhulme Trust

Specifying the GP mean

- Use the mean function to specify any functional form known to exist

- In the absence of prior information we tend to specify it as constant or a linear regression formula

$$h(x) = 1$$

$$h(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

The Leverhulme Trust

What does the prior mean function do?

UNIVERSITY OF LEEDS

- When there are very few simulator runs to give information the function is weighted towards the prior mean

- Outside the bounds of the simulator runs the emulator tends towards the prior mean

- With more (well-chosen) simulator runs the hyperparameters β are better estimated

The Leverhulme Trust

β hyperparameters

UNIVERSITY OF LEEDS

- Can be plugged in if a particular regression model is required

- Usually use maximum likelihood to estimate them and ‘plug-in’

- Could use a Bayesian approach but time-consuming

The Leverhulme Trust

Specifying the GP covariance

UNIVERSITY OF LEEDS

- The covariance function should reflect $c(\boldsymbol{x}, \boldsymbol{x}) = \mathbf{1}$ and that $c(\boldsymbol{x}, \boldsymbol{x}')$ decreases as $|\boldsymbol{x} - \boldsymbol{x}'|$ increases

- Includes further hyperparameters δ that specify the speed of covariance decrease with increasing $|\boldsymbol{x} - \boldsymbol{x}'|$

$$cov\{f(\boldsymbol{x}), f(\boldsymbol{x}')|\sigma\} = \sigma^2 \, c(\boldsymbol{x}, \boldsymbol{x}')$$

The Leverhulme Trust

Typical choices for $c(x, x')$

UNIVERSITY OF LEEDS

Gaussian:

$$c(x, x')|\delta = \exp\{-\frac{(x - x')^2}{2\delta^2}\}$$

Matern 5/2:

$$c(x, x')|\delta =$$

$$\left(1 + \frac{\sqrt{5}|x-x'|}{\delta} + \frac{5(x-x')^2}{3\delta^2}\right) exp(-\frac{\sqrt{5}|x-x'|}{\delta})$$

The Leverhulme Trust

Effect of covariance

UNIVERSITY OF LEEDS

- The uncertainty at simulator runs is 0
  - unless a nugget is used

- The uncertainty between points decreases as it gets closer to simulator suns

- The width of bounds between points increases as δ decreases

The Leverhulme Trust

The posterior distribution

- A Gaussian process, with parameters estimated by the training data

- Any point can be estimated

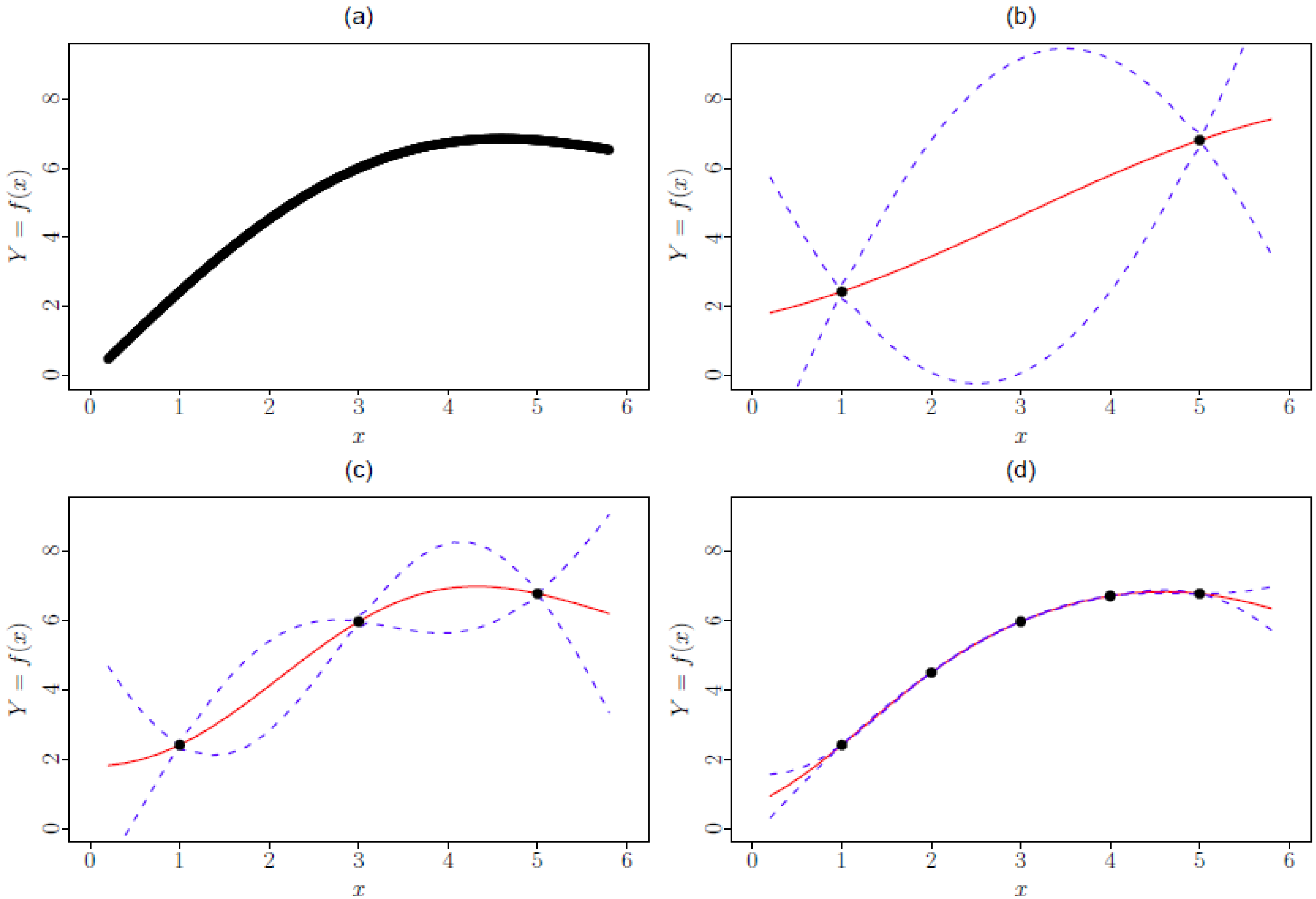
- Uncertainty in the estimate can also be estimated

Emulator uniqueness

- The emulator is non-unique
- Hyperparameters will be different if you re-run
  - Unless you set the seed
- Using prior information can help stability
- Often, it's not actually a problem…

Building an emulator

UNIVERSITY OF LEEDS

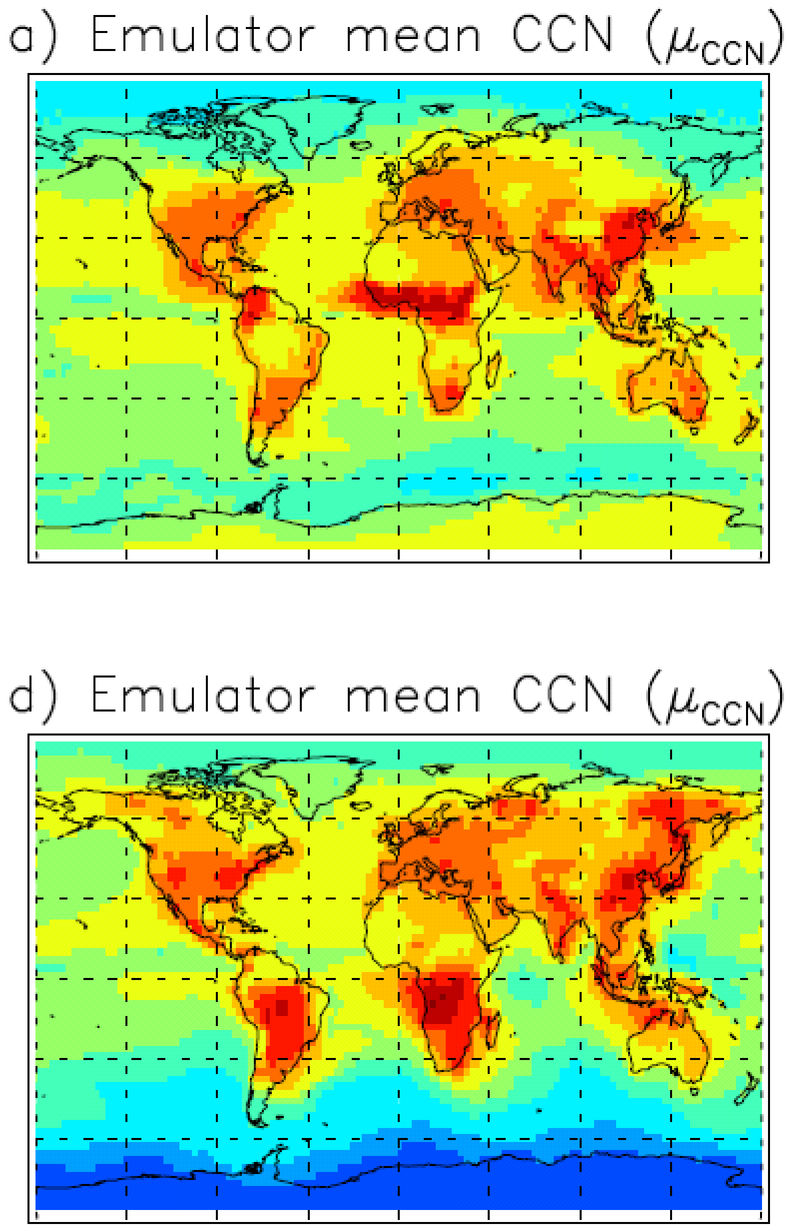


From Jill Johnson following O’Hagan (2006)

The Leverhulme Trust

Aerosol model emulation with DiceKriging (Roustant et al., 2012 & Lee,at al, 2013)

UNIVERSITY OF LEEDS



Emulation done gridbox by gridbox for sensitivity analysis

The Leverhulme Trust

Training data

UNIVERSITY OF LEEDS

- Need to update the prior with simulator points
- Require these points to provide good information about the function through uncertain space
- Often consider no prior information regarding particular parts of the space

The Leverhulme Trust

Design for training data
(McKay et al., 1979)

UNIVERSITY OF LEEDS

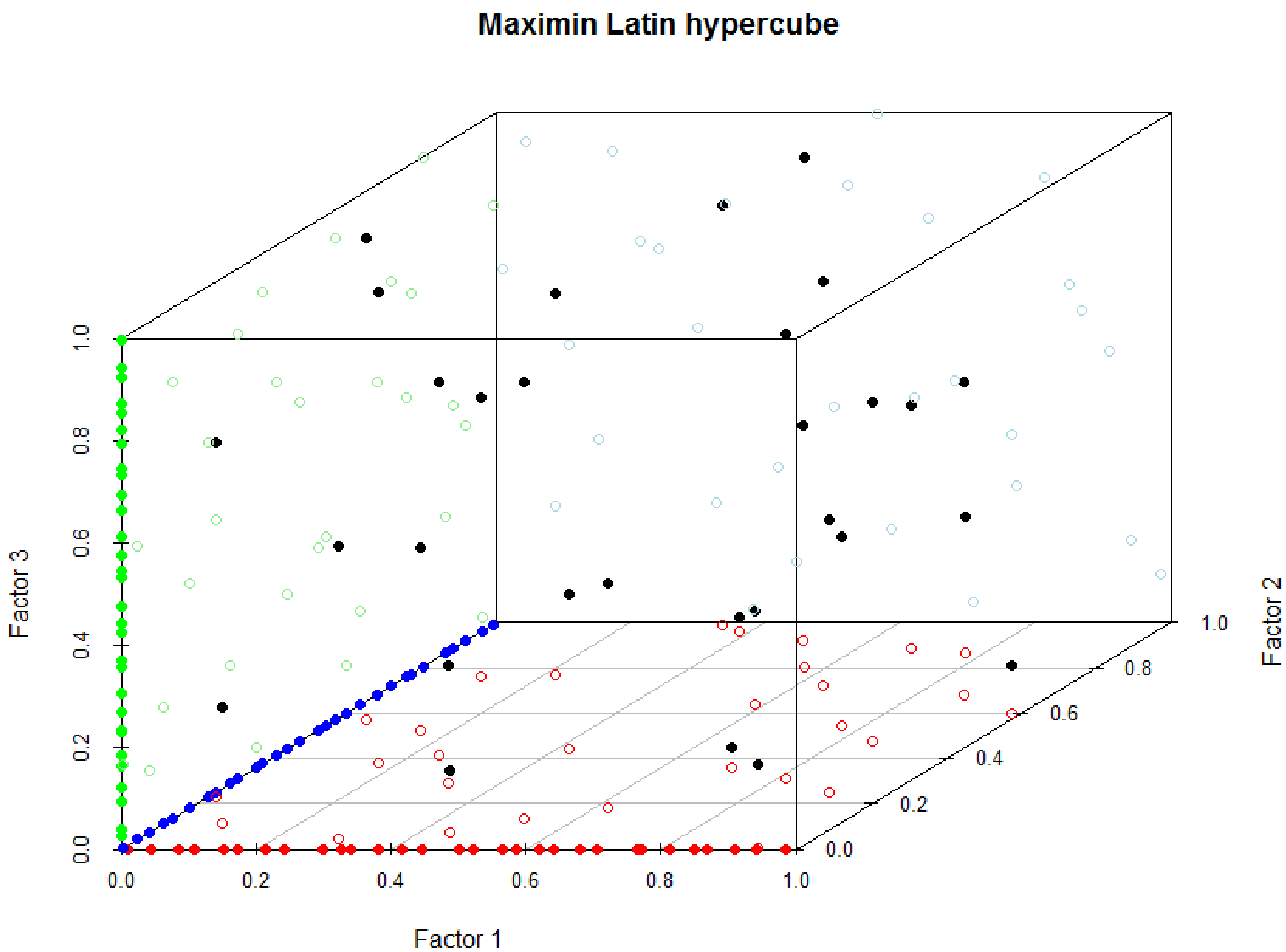
Latin hypercube sampling

- Good for inactive dimensions
- Use optimum or maximin
- May want piecewise LHS

The Leverhulme Trust

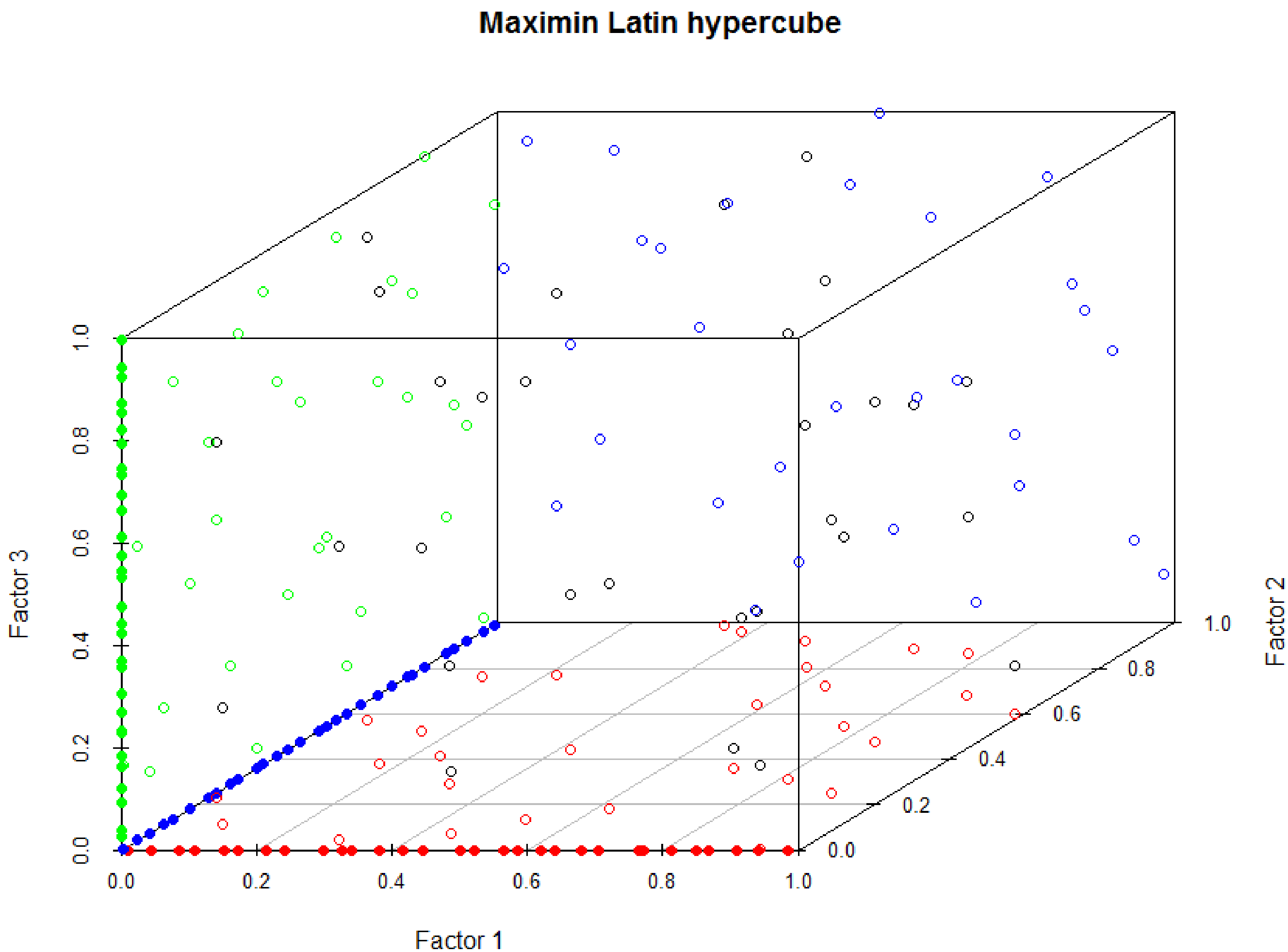
Latin hypercube sampling

UNIVERSITY OF LEEDS



The Leverhulme Trust

The marginals



The Leverhulme Trust

Sequential design

- When emulator validation suggests can update the design sequentially
- If no reason to search particular region, augment the design
- Can add points sequentially, perhaps Sobol, to particular regions

The Leverhulme Trust

Factorial designs

UNIVERSITY OF LEEDS

- These are not recommended
- The covariance function is better estimated using a variety of distances between points
- Uncertainty between training points will be very high

The Leverhulme Trust

Validation of emulators

- Have to check the emulator can predict what the emulator would say

- Usually want to check ‘out of sample’

- Can use leave one out when circumstances dictate

Aims of validation

UNIVERSITY OF LEEDS

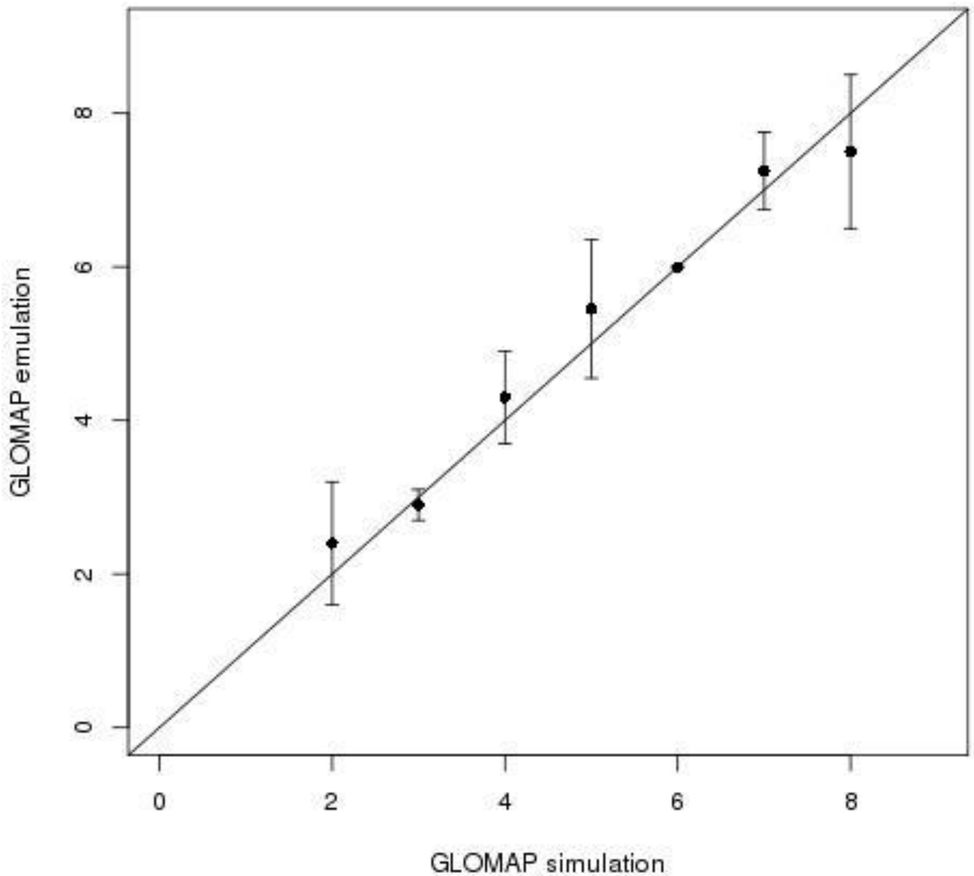
Prediction value

Ensure emulator predictions are close to simulator

Prediction uncertainty

Ensure uncertainty in emulator is small enough for further inference

The Leverhulme Trust


GLOMAP emulation
0
2
4
6
8
0
2
4
6
8
GLOMAP simulation

Validation design

- When out-of-sample tend to design two groups, as per Bastos & O'Hagan (2009)

- 1/3 close to simulated points and 2/3 spaced out
  - Helps reveal particular difficulties

Validation statistics

UNIVERSITY OF LEEDS

- Individual prediction errors
  - Treat like regression errors, approximately Gaussian and <|2| standardised
- Mahalanobis distance
  - single summary of individual prediction errors
  - extreme values indicate emulator/simulator mismatch
- Pivoted Cholesky errors
  - following variance decomposition
  - particular patterns aid interpretation of mismatch

The Leverhulme Trust

Validation plots

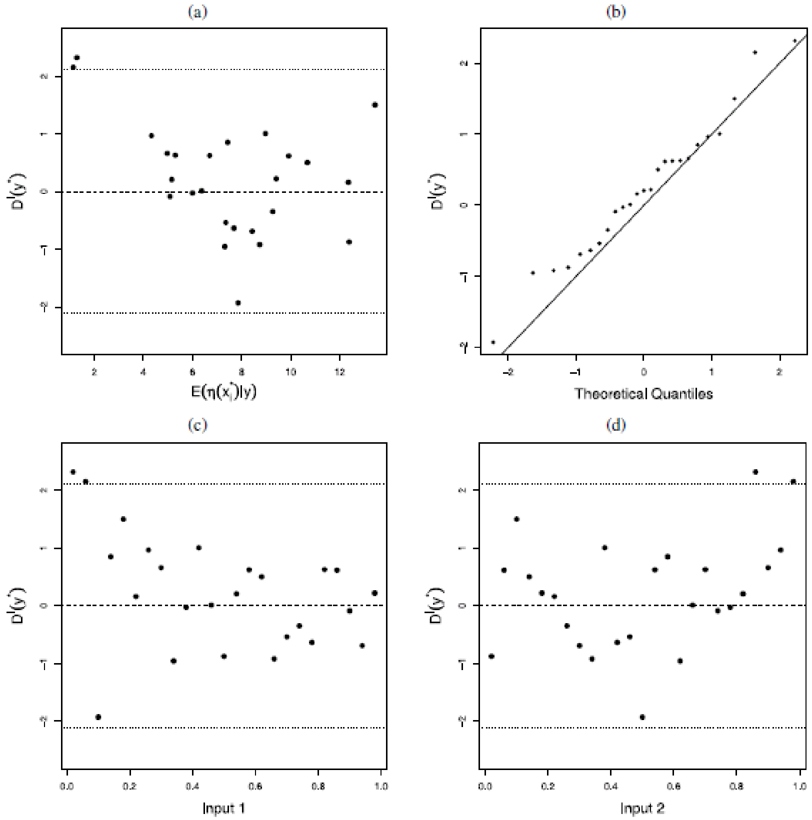
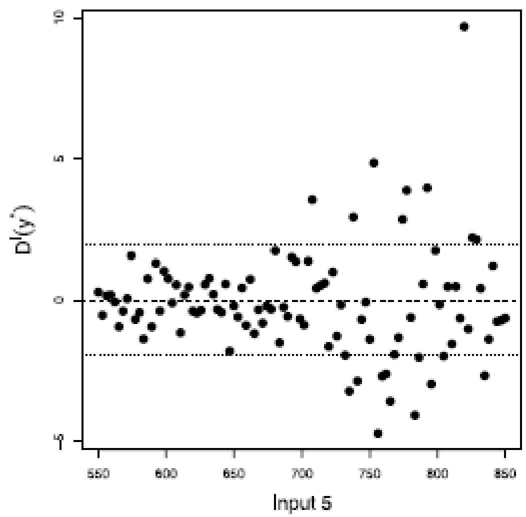
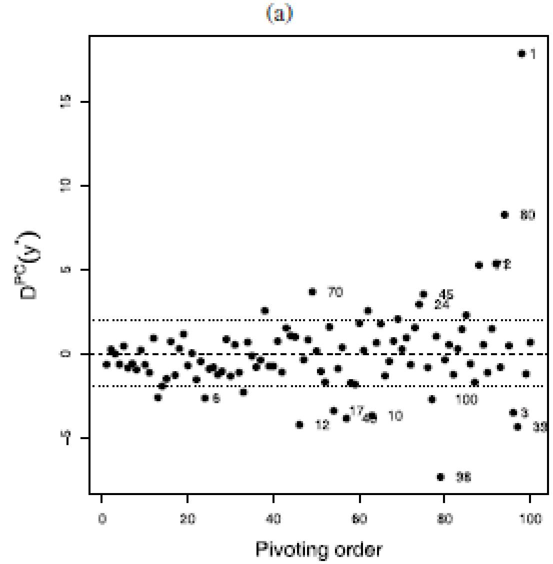


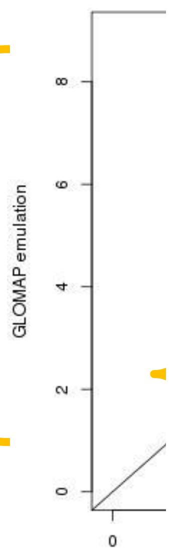
Figure 1. Graphical diagnostics for the toy example using the individual standardized errors: (a) $D^I(\mathbf{y}^*)$ against the emulator predictions; (b) QQ-plot; (c) $D_i(\mathbf{y}^*)$ against input 1; and (d) $D_i(\mathbf{y}^*)$ against input 2.

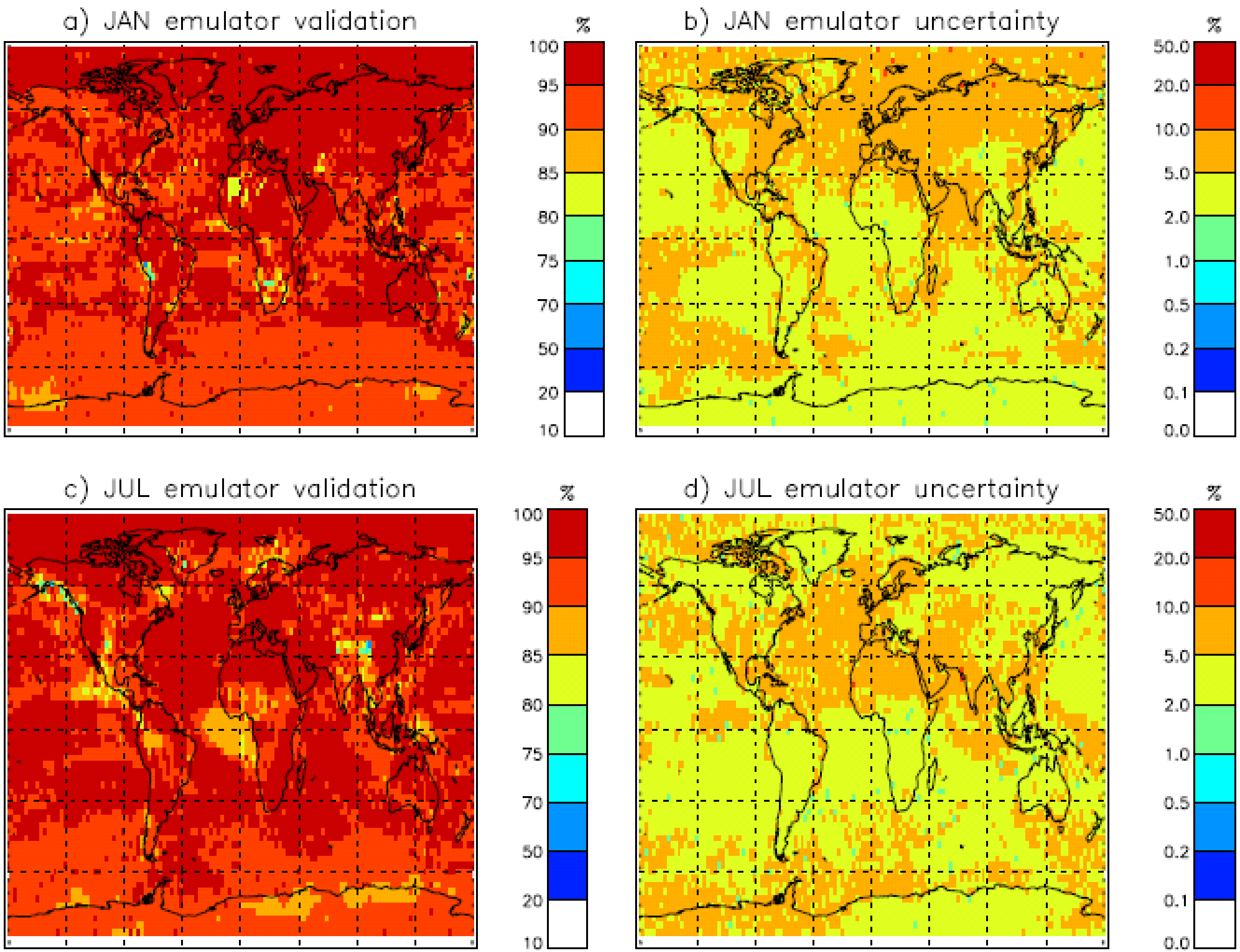


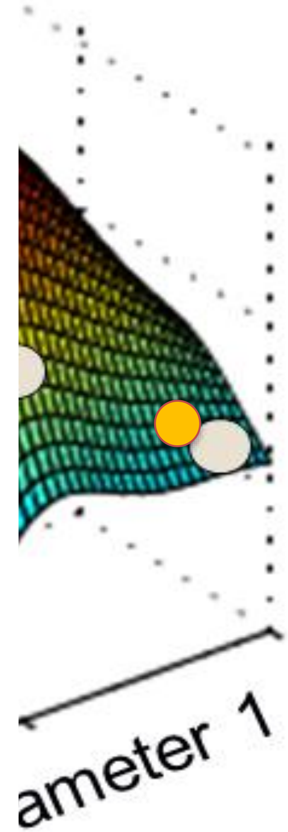


The Leverhulme Trust

Aerosol emulator validation







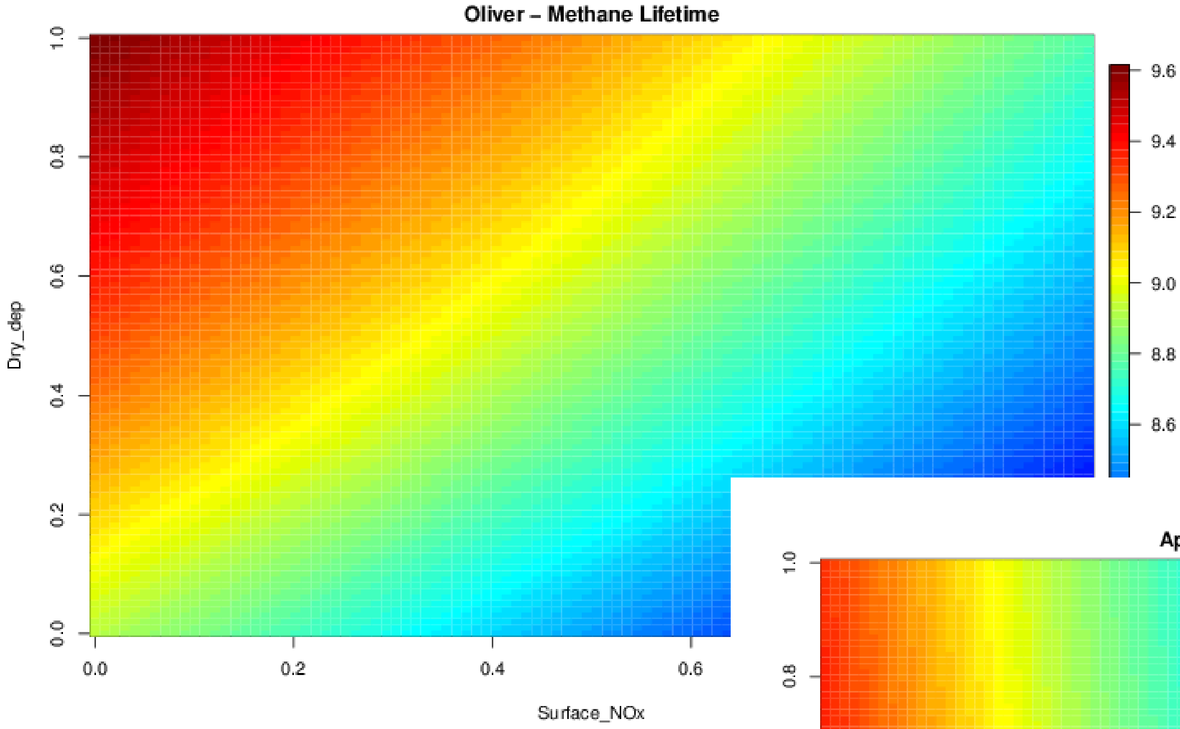
The Leverhulme Trust

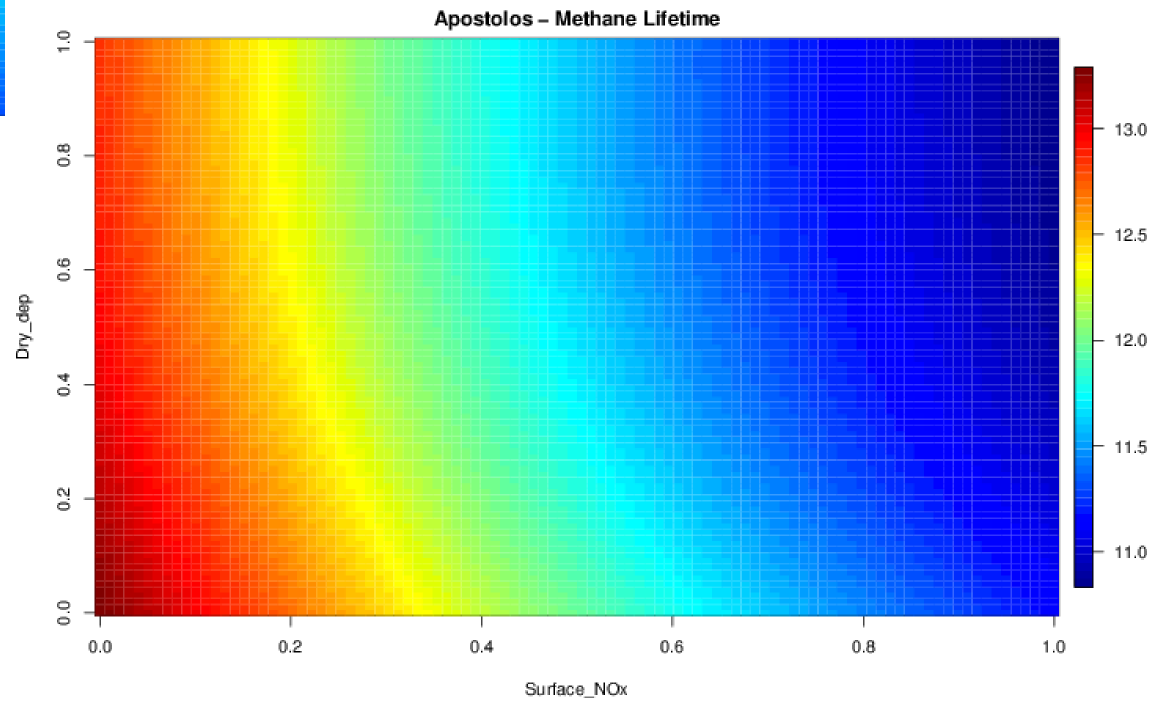
Leave one out

- When out-of-sample is not possible
- Leave a point out, build emulator, produce validation statistics
- Repeat for all points and compare statistics
- Any points with extreme statistics indicate problems

Emulation for understanding

UNIVERSITY OF LEEDS

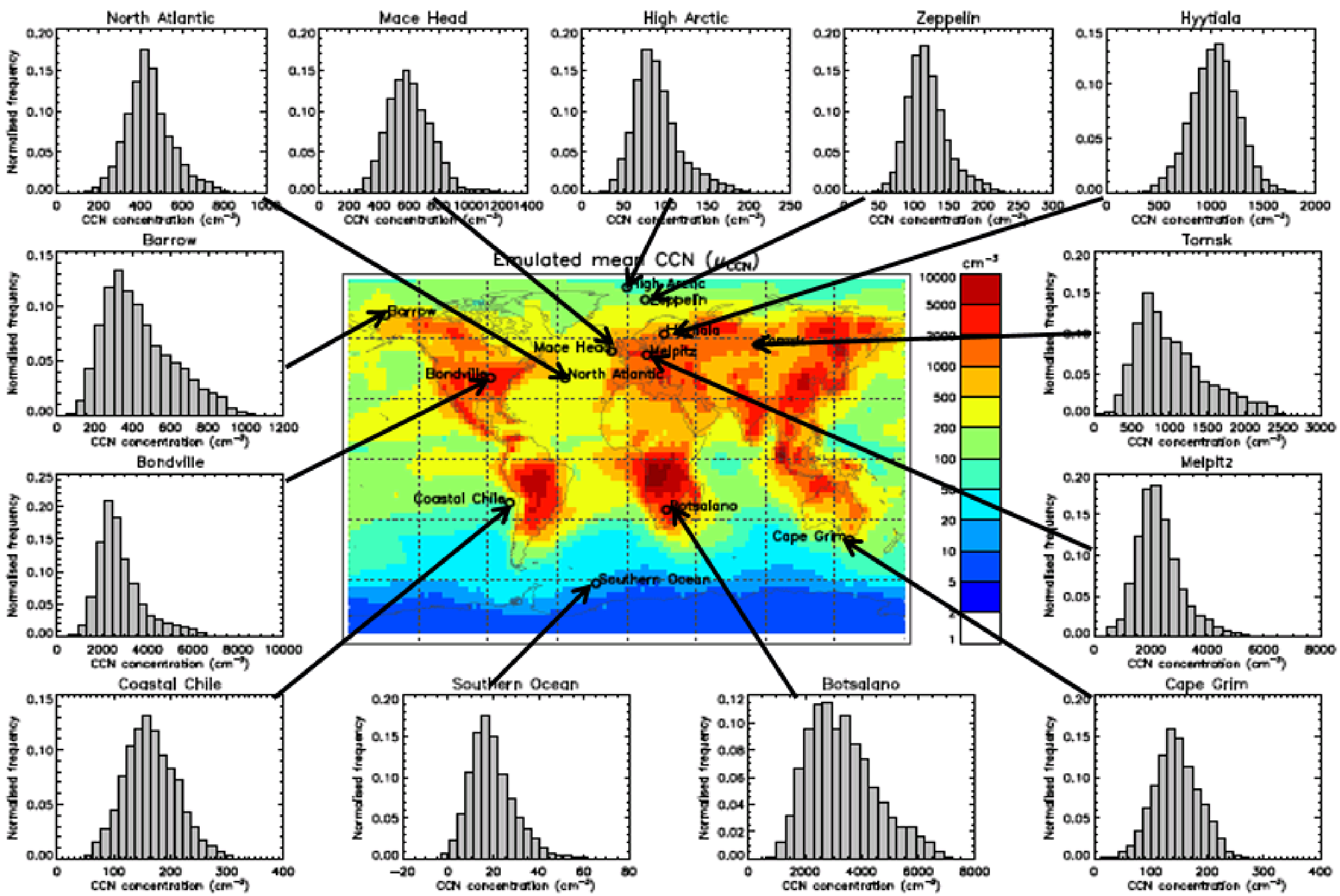




The Leverhulme Trust

Emulation for uncertainty analysis

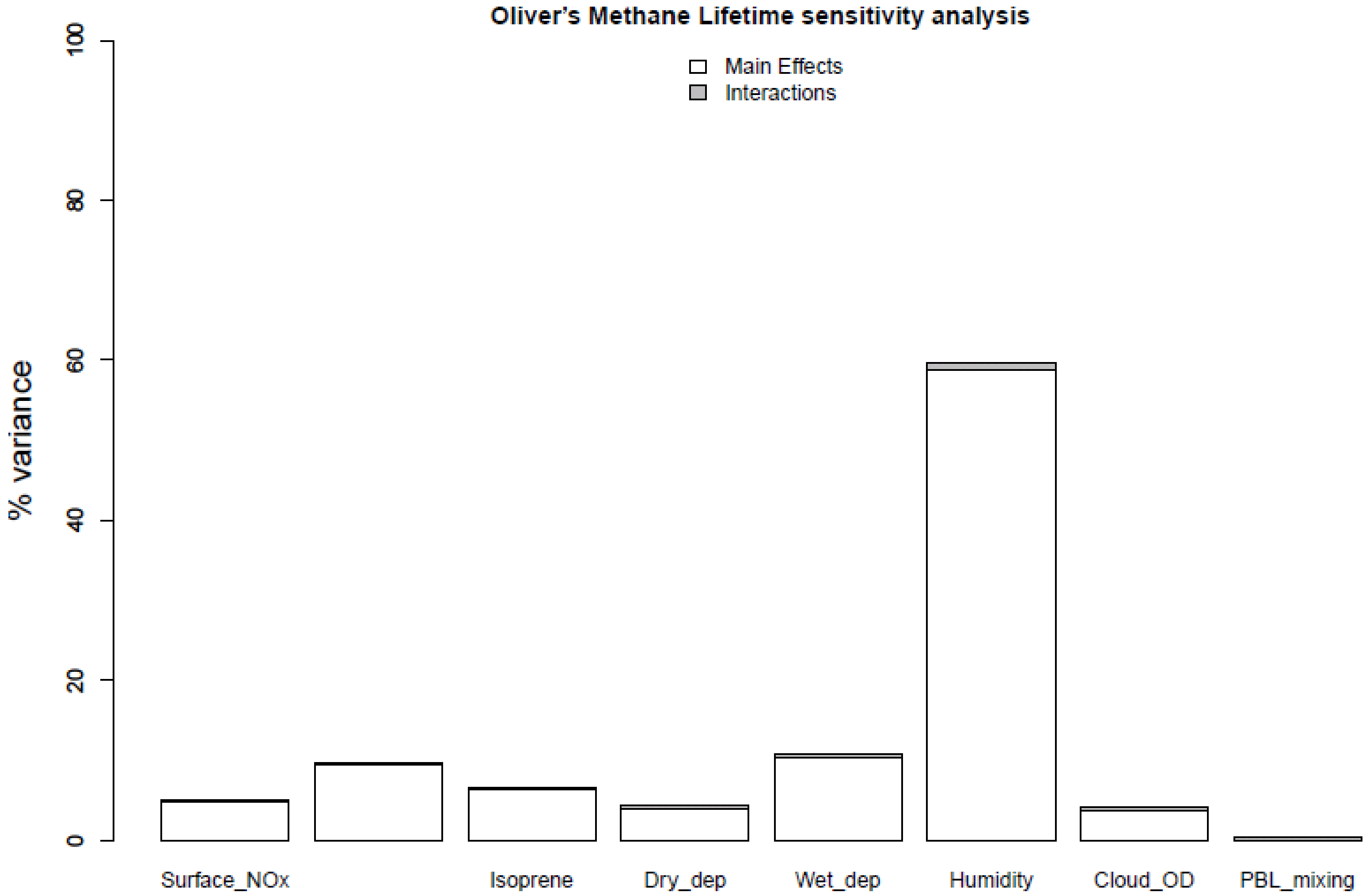
UNIVERSITY OF LEEDS



The Leverhulme Trust

Emulation for sensitivity analysis (Saltelli et al., 2000) using R sensitivity (Pujol et al., 2008)

UNIVERSITY OF LEEDS

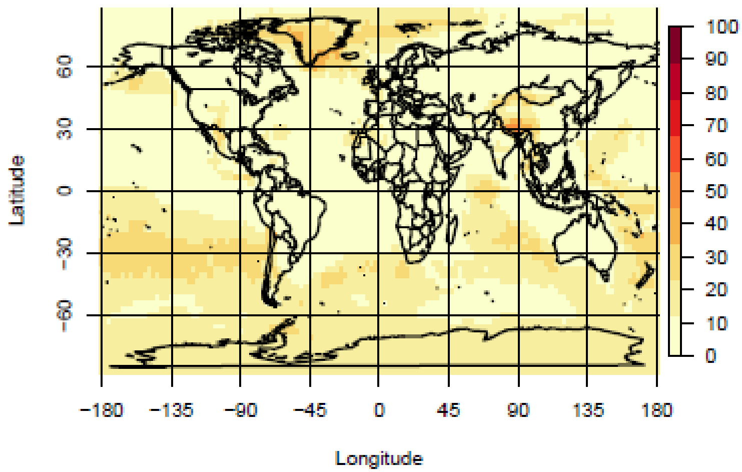


The Leverhulme Trust

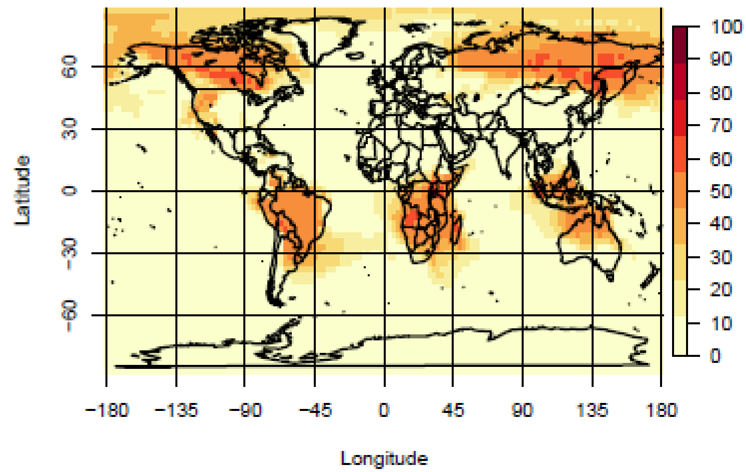
Aerosol model sensitivity analysis

UNIVERSITY OF LEEDS

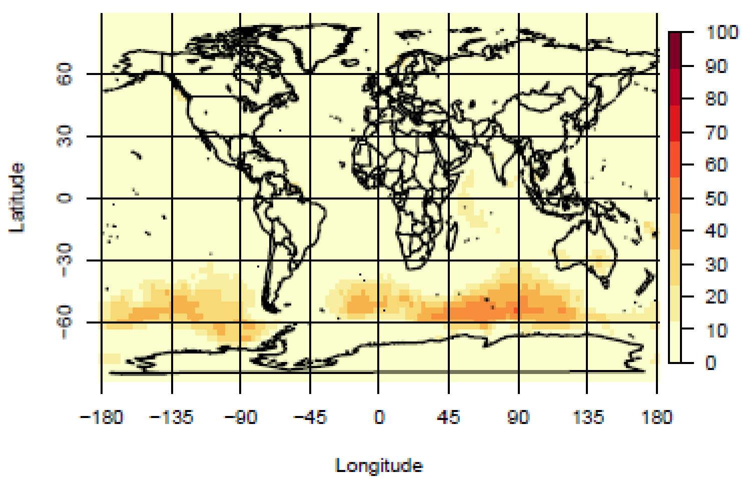
AIT_WIDTH JULY



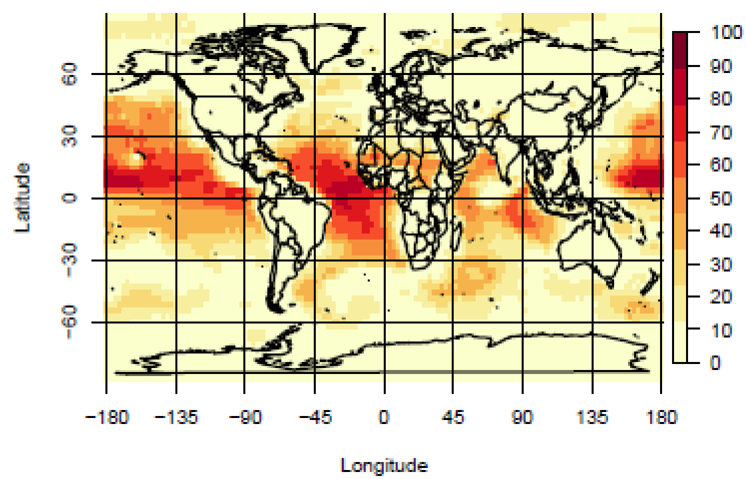
BB_DIAM JULY



SS_ACC JULY



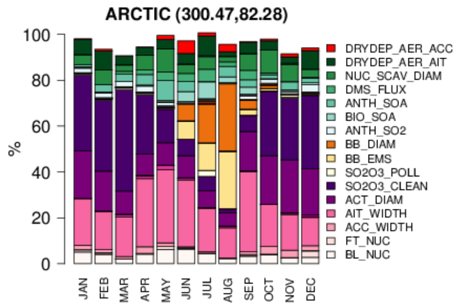
DRYDEP_AER_ACC JULY

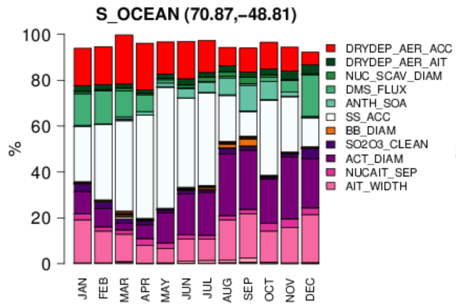


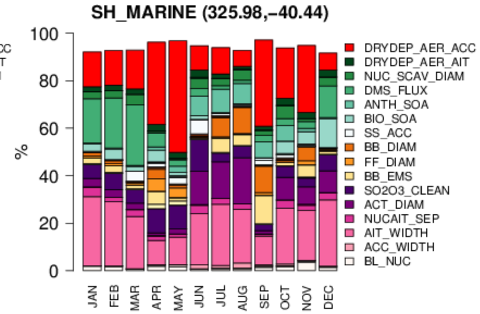
The Leverhulme Trust

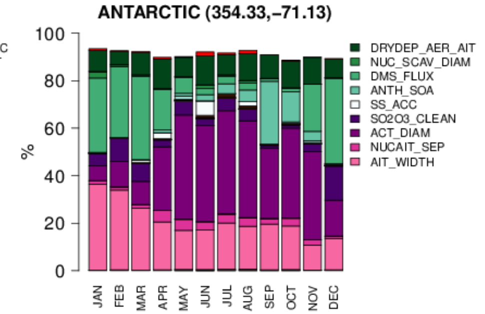
Aerosol model sensitivity analysis II

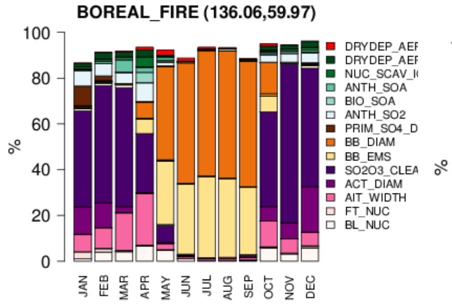
UNIVERSITY OF LEEDS

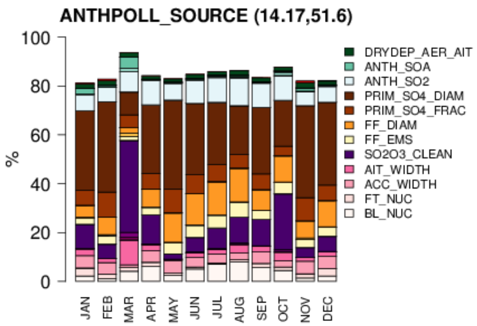


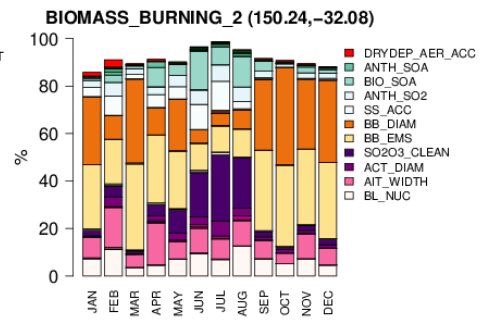


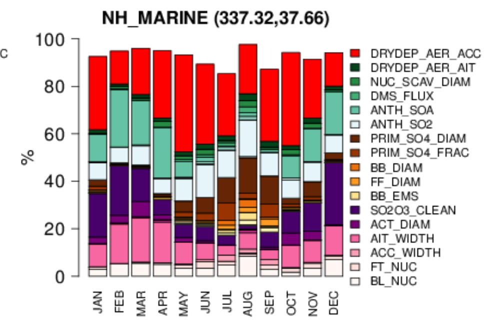


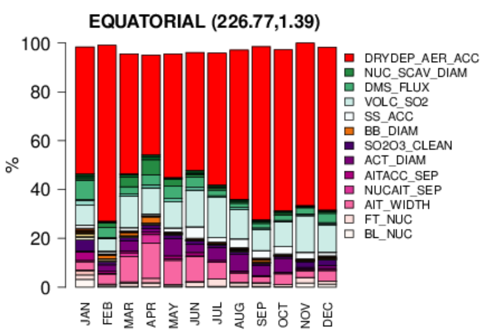


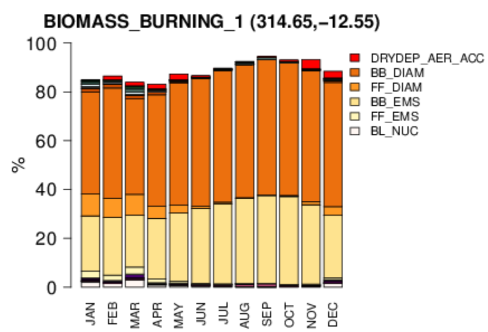






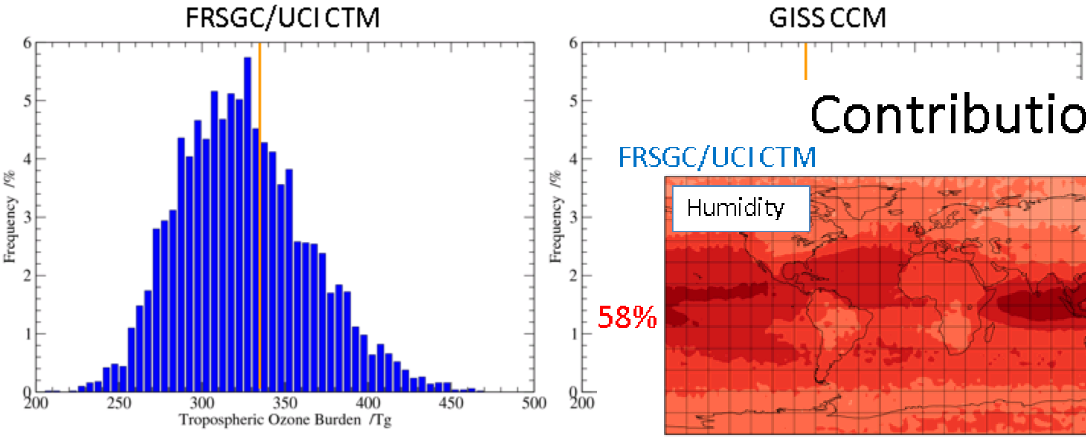




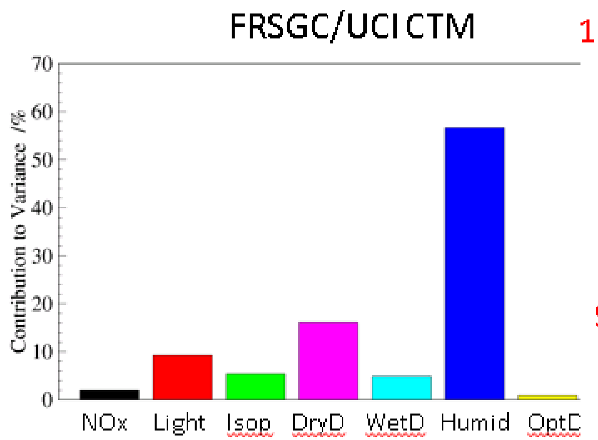


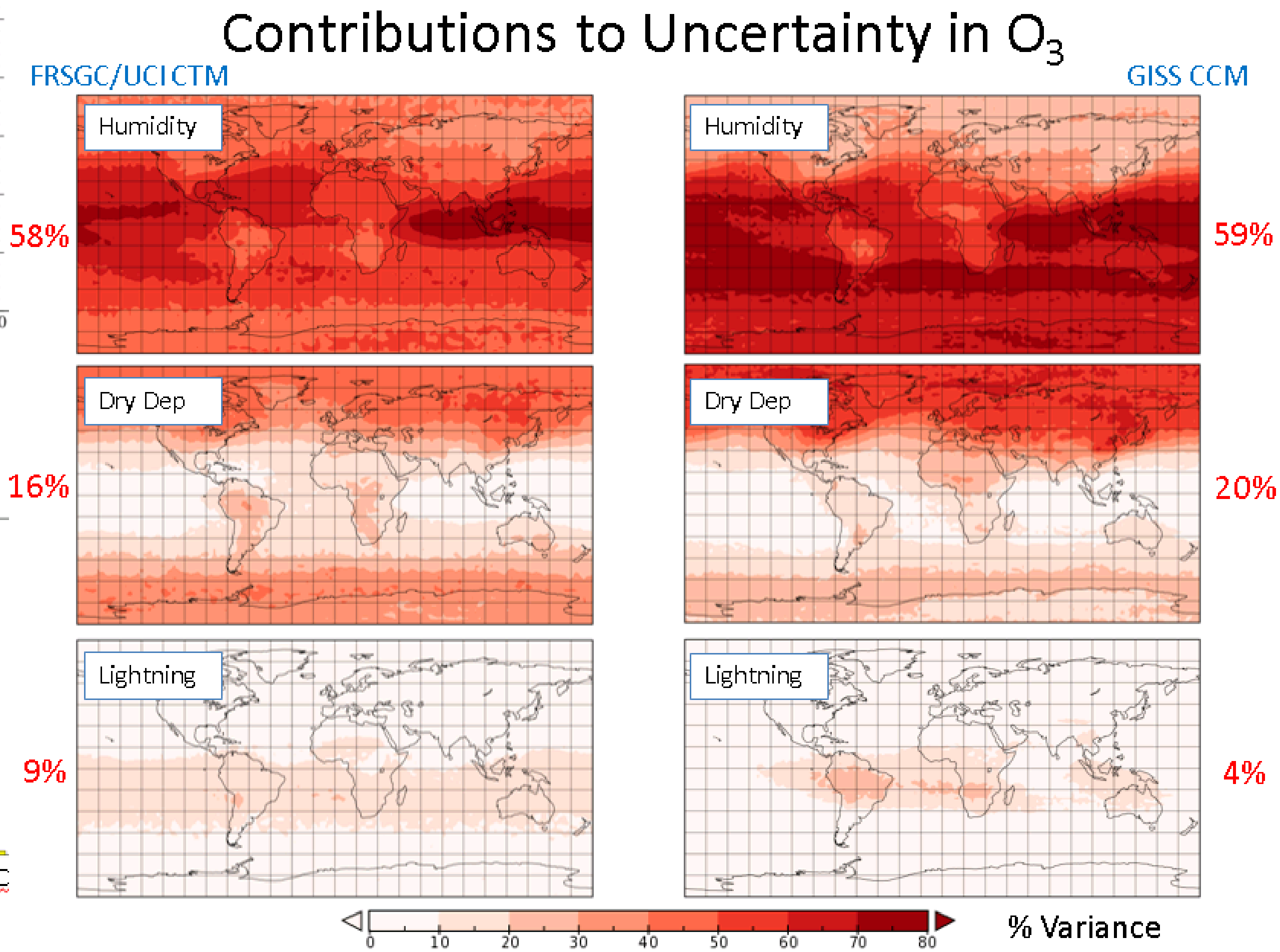
The Leverhulme Trust

Multiple model sensitivity analysis I



FRSGC: 325 ± 39 Tg





The Leverhulme Trust

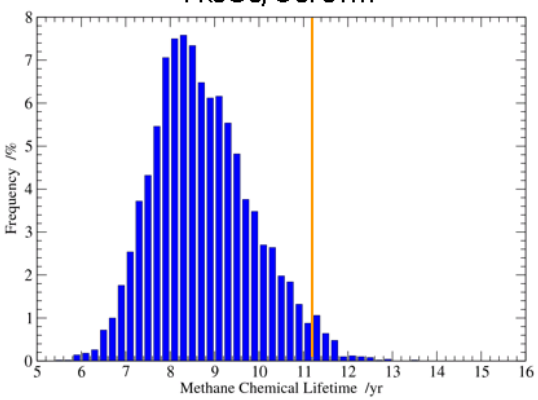
Multiple model sensitivity analysis II

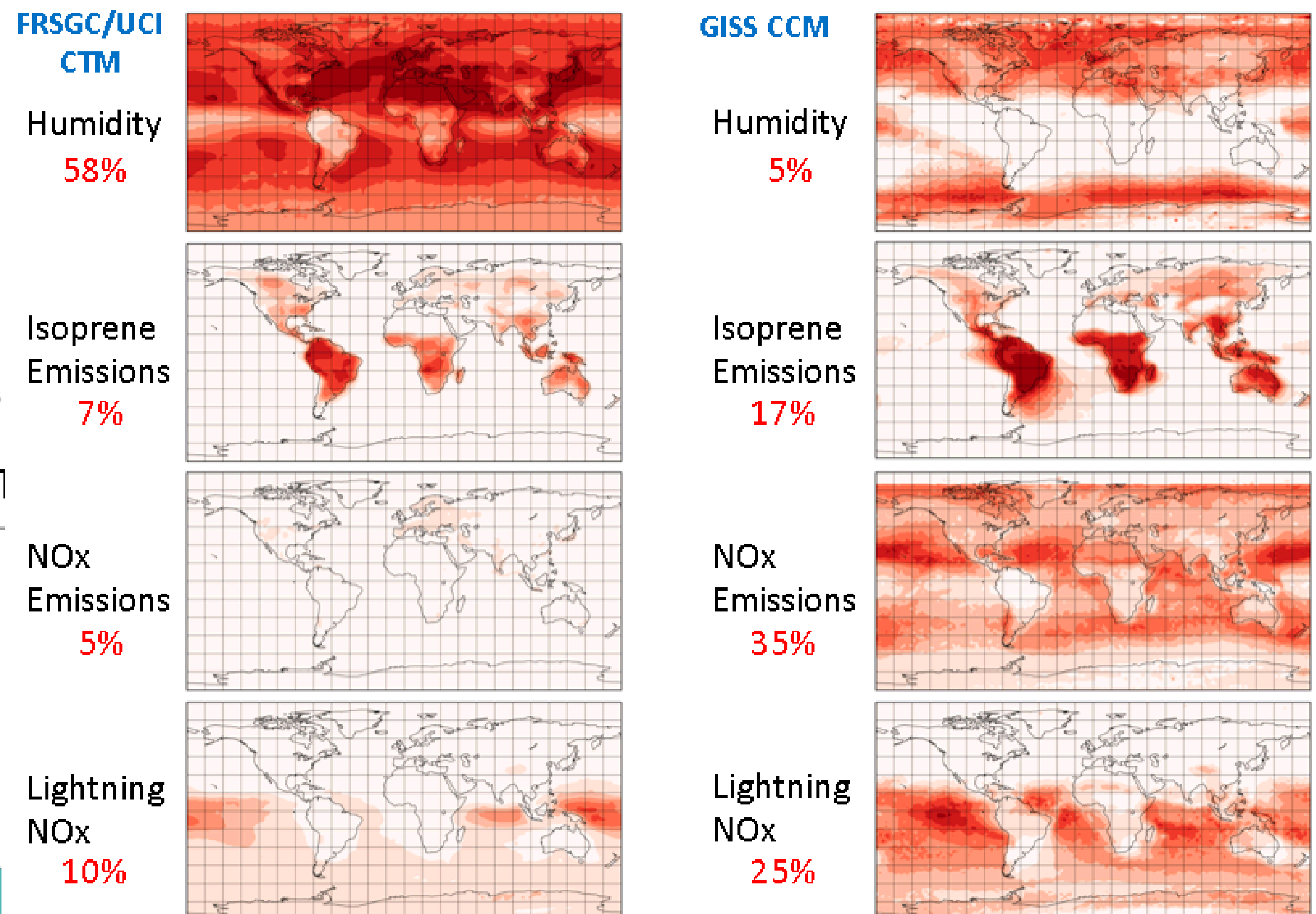
UNIVERSITY OF LEEDS

Tropospheric

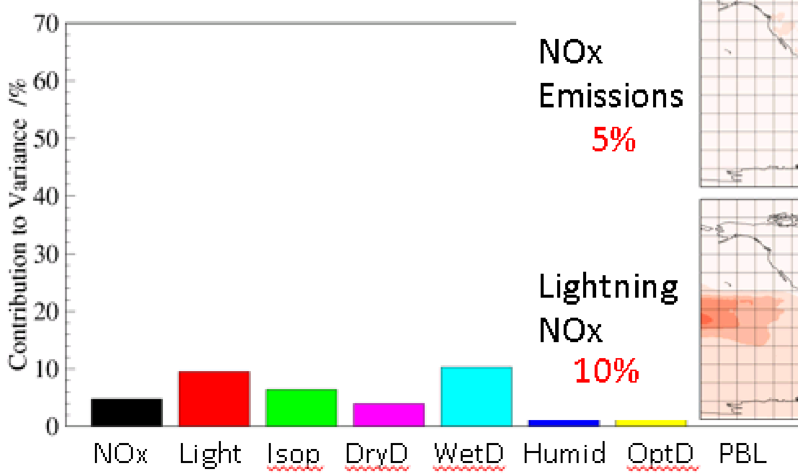
Contributions to Uncertainty in CH_4 Loss Rate

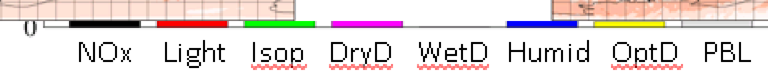
FRSGC/UCI CTM





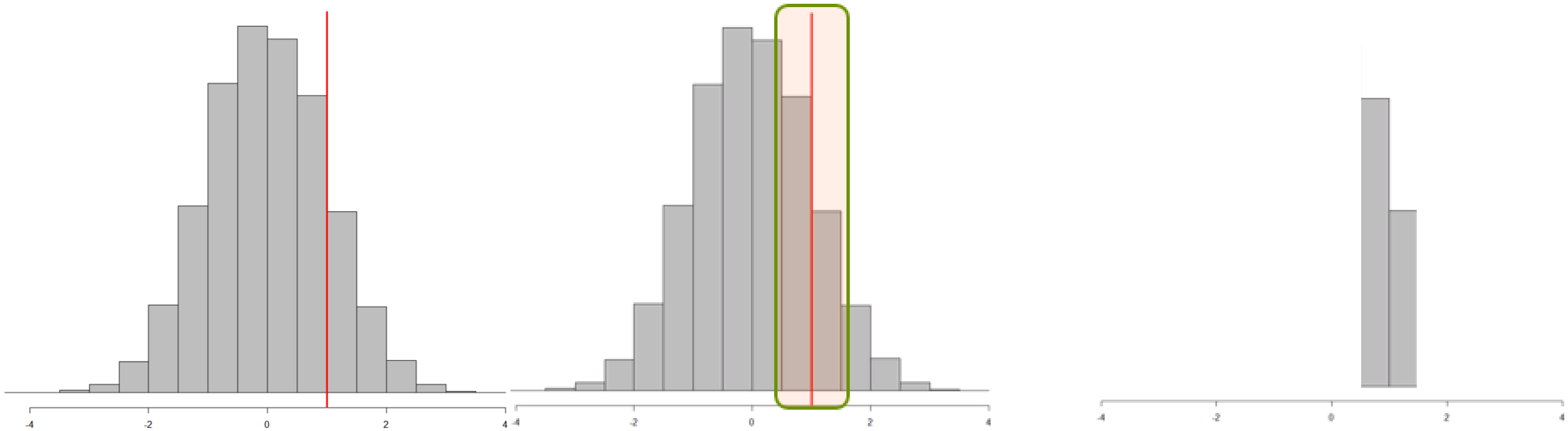
FRSGC/UCI CT

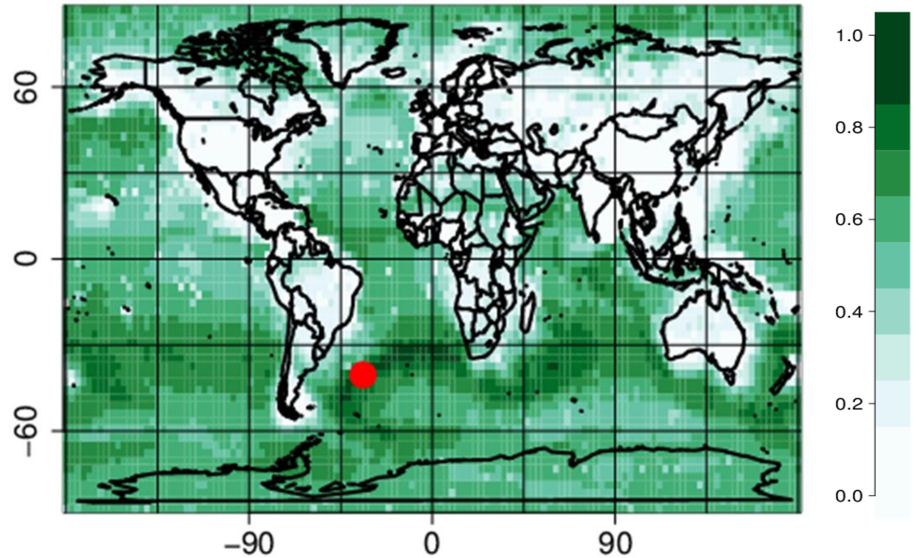




The Leverhulme Trust

Emulators for model constraint I

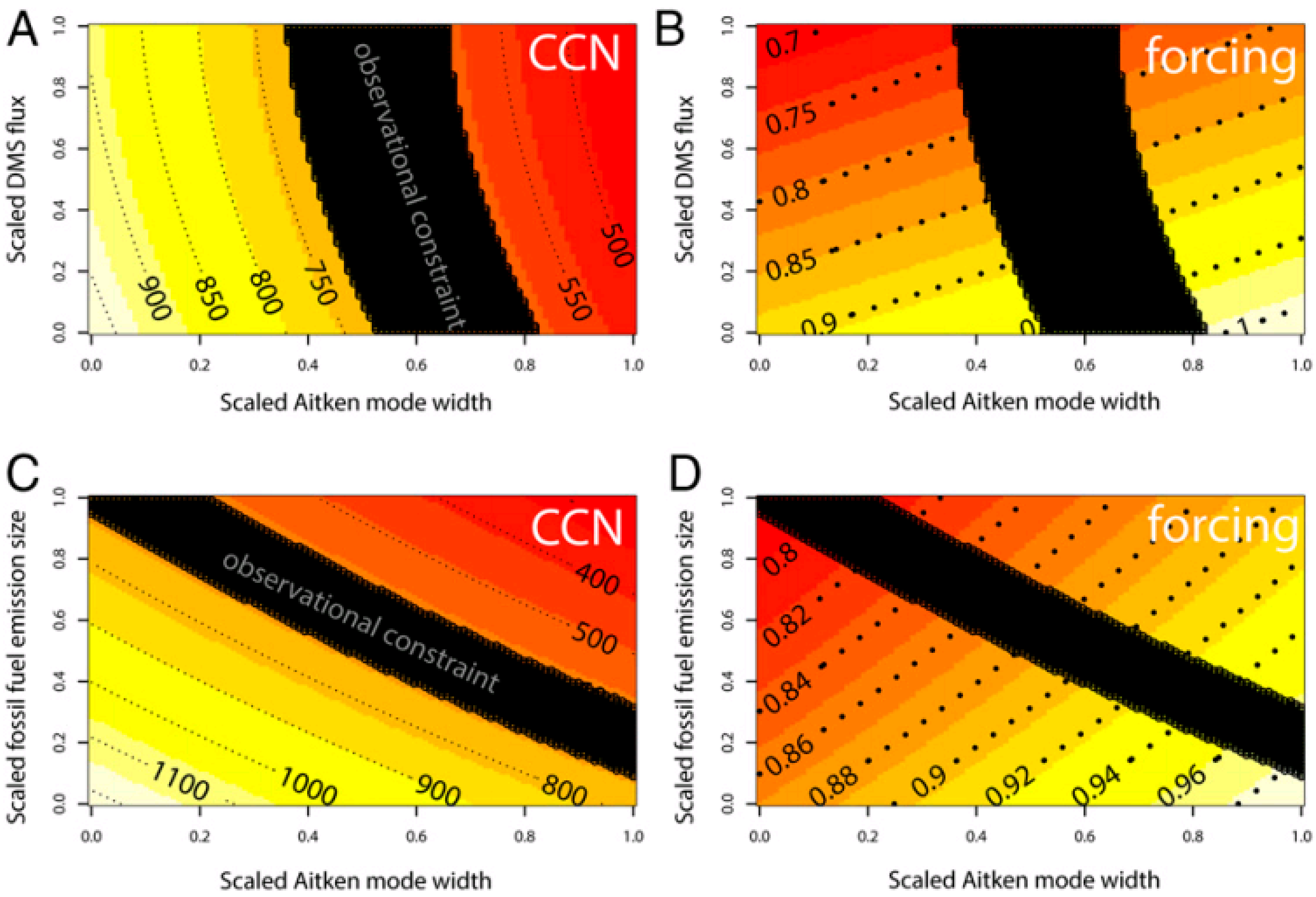




The Leverhulme Trust

Emulators for model constraint II

UNIVERSITY OF LEEDS



Extension to GP emulators

- Stochastic simulator – add an estimated nugget to covariance
- Multivariate emulator – must characterise the covariance between multiple outputs
- Discontinuities – could build multiple emulators, perhaps multivariate

What an emulator can't do

- Extrapolate too far with confidence
- Estimate a simulator output it's not trained to
- Replace the simulator

When not to use an emulator

UNIVERSITY OF LEEDS

- When the simulator is very cheap to run
- When it won’t validate
- When you haven’t got a good set of simulator runs for training
- When you aren’t sure what you’ll use it for

The Leverhulme Trust

Methodology references

UNIVERSITY OF LEEDS

- Rasmussen, Carl Edward, and Christopher KI Williams., 2006. Gaussian processes for machine learning. Vol. 1. Cambridge: MIT press.

- O'Hagan, A., 2006. Bayesian analysis of computer code outputs: a tutorial. Reliability Engineering & System Safety, 91(10), pp.1290-1300.

- Roustant, O., Ginsbourger, D. and Deville, Y., 2012. DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization.

- McKay, M.D., Beckman, R.J. and Conover, W.J., 1979. Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, 21(2), pp.239-245.

- Bastos, L.S. and O'Hagan, A., 2009. Diagnostics for Gaussian process emulators. Technometrics, 51(4), pp.425-438.

- Saltelli, A., Chan, K. and Scott, E.M. eds., 2000. Sensitivity analysis (Vol. 1). New York: Wiley.

- Pujol, G., Iooss, B. and Iooss, M.B., 2008. Package 'sensitivity'.

The Leverhulme Trust

Selected Lee references

UNIVERSITY OF LEEDS

Lee LA; Reddington CL; Carslaw KS (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**, pp.5820-5827. doi: 10.1073/pnas.1507050113

Carslaw KS; **Lee LA**; Reddington CL; Pringle KJ; Rap A; Forster PM; Mann GW; Spracklen DV; Woodhouse MT; Regayre LA; Pierce JR (2013) Large contribution of natural aerosols to uncertainty in indirect forcing, *Nature*, **503**, pp.67-71.

Lee LA; Pringle KJ; Reddington CL; Mann GW; Stier P; Spracklen DV; Pierce JR; Carslaw KS (2013) The magnitude and causes of uncertainty in global model simulations of cloud condensation nuclei, *Atmospheric Chemistry and Physics*, **13**, pp.8879-8914. doi: 10.5194/acp-13-8879-2013

Lee LA; Carslaw KS; Pringle KJ; Mann GW (2012) Mapping the uncertainty in global CCN using emulation, *Atmospheric Chemistry and Physics*, **12**, pp.9739-9751. doi: 10.5194/acp-12-9739-2012

Lee LA; Carslaw KS; Pringle KJ; Mann GW; Spracklen DV (2011) Emulation of a complex global aerosol model to quantify sensitivity to uncertain parameters, *ATMOSPHERIC CHEMISTRY AND PHYSICS*, **11**, pp.12253-12273. doi: 10.5194/acp-11-12253-2011

The Leverhulme Trust