All Models Are Wrong

Tamsin Edwards

University of Bristol

NCAS Summer School

September 2013

Introductions: you

What models are you using?

Introductions: this lecture

I will assume:

- Your climate model is computationally expensive
- Your climate model is deterministic
- You care about uncertainty
- You want your uncertainty assessments to be meaningful

All models are wrong

"All models are wrong, but some are useful"
– George Box (1919-2013)

- George meant statistical models, but we will apply this to simulators
- How wrong?
- What is our uncertainty about how wrong?

Why should we care about uncertainty?

Results are *meaningless* without context, i.e. without uncertainty estimates.

Some approaches to uncertainty:

- expert assessment: "From knowledge of the simulator (*optional: and comparing with observations*) we judge the simulator to have these biases."
- informal sensitivity study: "We tried various modelling choices and got these results (*optional: this set of choices gives results most similar to observations*)."

Why should we care about uncertainty?

(cont.)

- history matching: "We systematically sampled various modelling choices, and after comparing with observations, x results were not ruled out (*optional: according to a 95% confidence interval*)."
- Bayesian calibration: "We systematically sampled our uncertainties about various modelling choices, and after comparing with observations, our probabilistic assessment (*distribution, 95% credibility interval, maximum probability etc*) for the results is this."

The last two are statistically-founded methods; this lecture will give an overview of them.

Why make statistical assessments?

The field of statistics is the science of uncertainty. It provides a universal framework within which different studies can be interpreted and compared.

Example: recent sea level projections from ice2sea.

Why make statistical assessments?

We presented a range of results from different plausible modelling options, i.e. an informal sensitivity study.

> *"We should be cautious of comparing with AR4...theirs are stated as a 5-95% range while ours don't have a probability range attached."*

Everyone: So how should we compare the new results to the old?

Us: Er...

Statistically-founded assessments of uncertainty

The challenge is to assess the limitations of the simulator within a statistical framework; to translate "There is a lot of uncertainty in climate modelling..." into formal statistical inference.

This inference may be in the form of a probabilistic statement such as a confidence interval (in HM) or credibility interval (in BC), or a non-probabilistic - but still statistically-founded - statement about implausible modelling choices (in HM).

This is a very complicated business :(and you should consider consulting a statistician!

Sources of simulator uncertainty

- Initial conditions. Our uncertainty about the state of the system at the start of the simulation.
- Boundary conditions. Our uncertainty about the boundary conditions that constrain the system evolution.
- Parametric. Our uncertainty about the best values of the coefficients in the simulator.
- Structural. Our uncertainty about the difference between the simulator output and the system, even were we to know the best values of the parameters.

We will focus on 'internal' simulator uncertainties, i.e. parametric and structural.

Setting the scene

We can describe the simulator as:

$$\mathrm{sim}(s_i; x, \theta) \quad i = 1, \ldots, n$$

where:

- sim Simulator.
- s_i Index of the simulator output: in general the variable, location, and time.
- x Initial and boundary conditions.
- θ Parameters.

Simulator parameters

A common assertion is that there exists a parameter value θ^* such that

$$\mathrm{system}(s_i; x) \approx \mathrm{sim}(s_i; x, \theta^*) \quad \text{for all } s_i \text{ and } x.$$

Our uncertainty about the value of θ^* is the parametric uncertainty.

Simulator discrepancy

But $\mathrm{system}(s_i; x)$ is *not* equal to $\mathrm{sim}(s_i; x, \theta^*)$...!

Conceptually, the relationship between the system and the simulator can be thought of as

$$\mathrm{system}(s_i; x) = \mathrm{sim}(s_i; x, \theta^*) + \mathrm{disc}(s_i; x)$$

where 'disc' is the discrepancy (i.e. difference) between the simulator at its 'best' parameterisation, θ^*, and the actual system values.

Our uncertainty about the value of disc is the discrepancy variance, a.k.a. structural uncertainty.

Inference about parameters

We have already mentioned two popular 'levels' of learning about θ from observations: history matching and Bayesian calibration.

Both aim to quantify and reduce parametric uncertainty in the presence of structural uncertainty.

In a deterministic simulator, we can directly translate assessments of uncertainty in parameters to uncertainty in outputs.

Inference about parameters

1. History matching

Ruling out unacceptably bad ('implausible') candidates for θ^*.

If we make careful choices for:

- how to measure model-data differences
- thresholds for ruling out

we can make meaningful statements about which values of the parameters are 'not ruled out yet'.

We might also be able/choose to express these as confidence intervals.

Inference about parameters

2. Bayesian calibration

Updating our judgements about the probability of different values of θ^*.

If we make careful choices for:

- our initial judgements about θ
- how to measure model-data differences

we can infer probability distributions, and, from these, credibility intervals, for the simulator parameters and output.

The two are complementary. We can do HM to reduce the parameter space before BC, or aim for full BC with the fall-back option of HM.

Confidence or credibility?

Confidence interval:

- a random range estimated to contain the value of an unknown fixed parameter
- "If I repeat this statistical analysis 100 times, 95 of the 95% CIs I obtain will contain the true value of θ^*."

Credibility interval (Bayesian):

- a fixed range estimated to contain most of the probability density of an unknown random parameter
- "I make this prior estimate of the distribution of θ^*, and update it with this set of observations, and the range that contains 95% of the resulting posterior probability density is this."

If you observe this distinction, you are doing better than the IPCC reports...

Overview

A Experimental Design. Systematically sample plausible modelling choices.

B Observational Comparison. Compare results with real world.

C Statistical Inference. Interpret results.

It is more efficient to proceed sequentially: A-B-A-B-...-C.

Section A: Experimental design

Section A: (Initial) Experimental Design

1. Choose uncertain quantities to sample
2. Express these as probability distributions, as far as possible
3. Sample efficiently

Section B: Observational Comparison

Section A: (Extension to) Experimental Design

4. Increase ensemble size efficiently

1. Choose uncertain quantities to sample

List your parameters in decreasing order of importance.

Choose the top few as the active parameters for your ensemble (ten is a lot).

It is often difficult to predict the most important parameters in advance.

So it is sensible to use only some of your computing budget on the initial ensemble, using it as a screening study of a larger number of parameters.

Save the rest of your budget for exploring the parameters you later find are most important.

Express uncertain quantities as probability distributions

For BC we must express uncertain quantities as probability distributions: the sum of probabilities over all choices must equal one.

So we must know all physically meaningful values of the uncertain quantity so that we can span the space and partition it without gaps or overlaps.

Do these qualify?

- Continuous simulator parameters?
- Discrete simulator parameters?
- Structural switches (choosing different schemes)?

For HM we must still express uncertain quantities as distributions to sample, but coverage and shape are less important.

Express other uncertain quantities as scenarios

If we do not know the values that fully span or properly partition the space, we can only use plausible scenarios.

Scenarios are typically used for boundary condition uncertainties too.

Summarising.
We can easily summarise probability distributions: for example, with mathematical expectation.

Scenarios are much more difficult to summarise and interpret. The safest thing to do is to present results for each scenario separately.

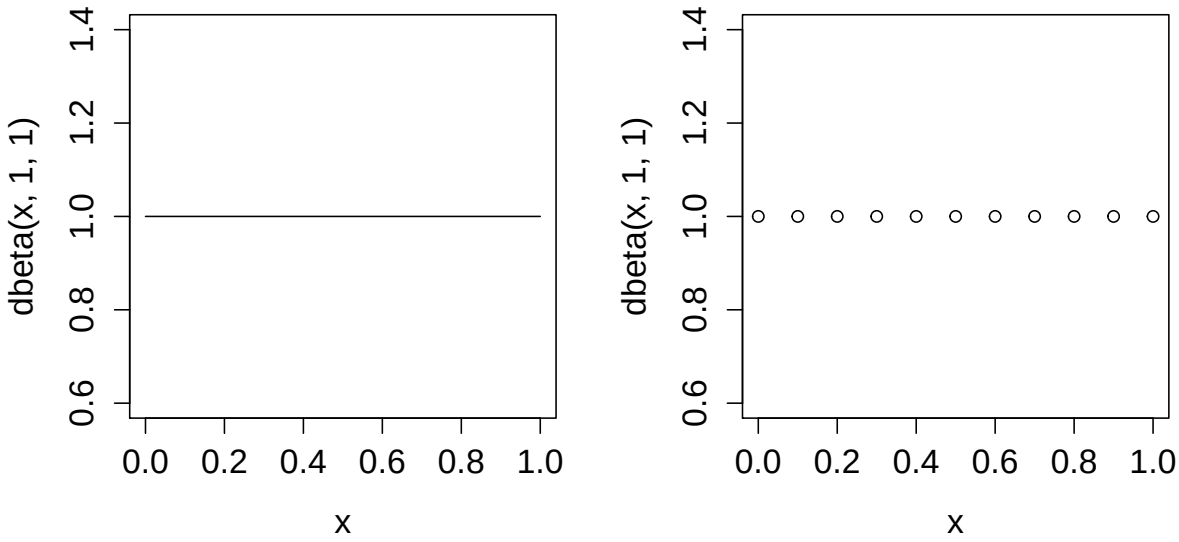
Q. How do MMEs fit into all this?

Choosing probability distributions

The most common way to set probability distributions is to:

1. elicit physically meaningful ranges from experts
2. assign equal probability to each value in that range

i.e. use uniform distributions (continuous or discrete).



Uniform distributions

It is common to use uniform distributions because it is:

- Easy
- Feels like it uses no judgement about probability
  - N.B. it does, though this is not a problem if you are aware of it

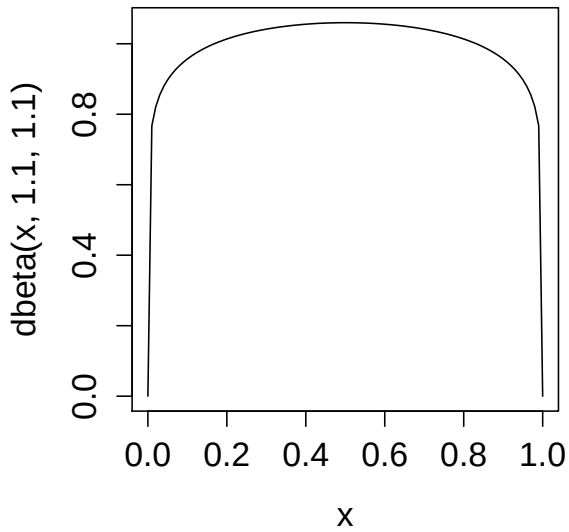
Consider expanding the ranges given to you by experts...

You might choose to transform the parameters to a [0,1] interval and use the 'standard uniform distribution': this is a special case of the beta distribution, Beta(1.0, 1.0).

Non-uniform distributions

We can use other distributions: directly, or by transforming the parameter and expressing this as a uniform distribution.

But do not use distributions that are too peaked (e.g. triangular): it will concentrate all the simulations into a very, very, very tiny volume. They should roll off only slightly at the ends of their ranges, e.g., Beta(1.1, 1.1):



Sampling efficiently

Vary all parameters at once, for the interactions.

You may have 'only' 10 parameters, but a 10-D space is very large. You would need 1000s of simulations to span it directly (e.g. min, central and max for all).

So you need a space-filling design. A Latin hypercube is very popular, because it is easy.

A Latin hypercube:

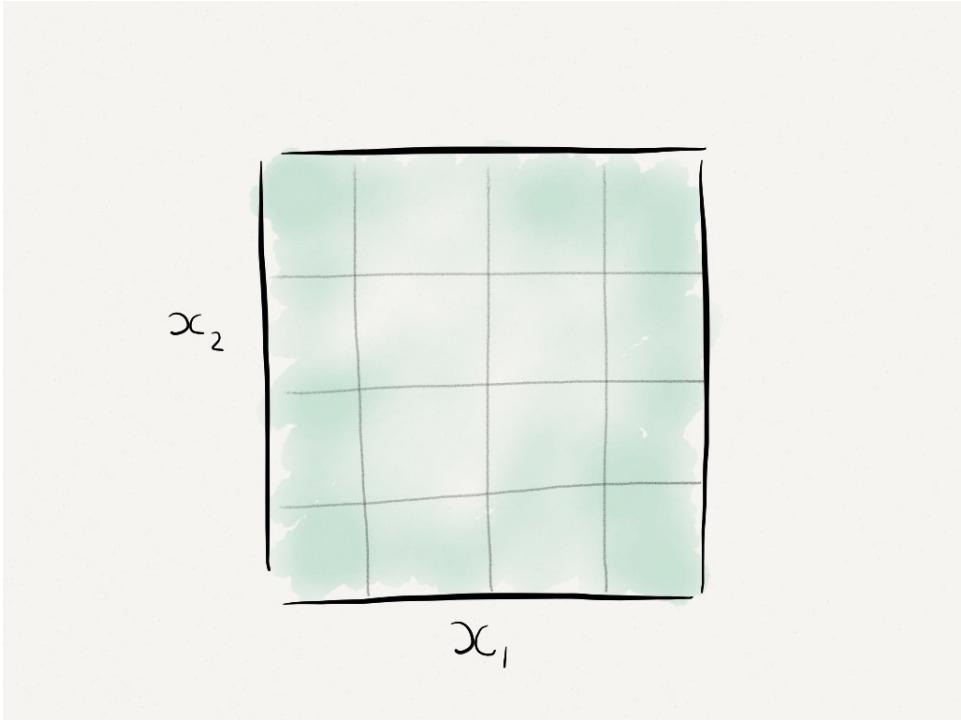
- partitions each parameter range into equally probable intervals
- generates (in 2D) one sample in each row and column; for higher dimensions, axis-aligned hyperplanes
- is designed to give a particular number of samples

Use a maximin Latin hypercube to maximise the minimum distance between points.

Maximin Latin Hypercube (Doug McNeall)

An aside on experiment design

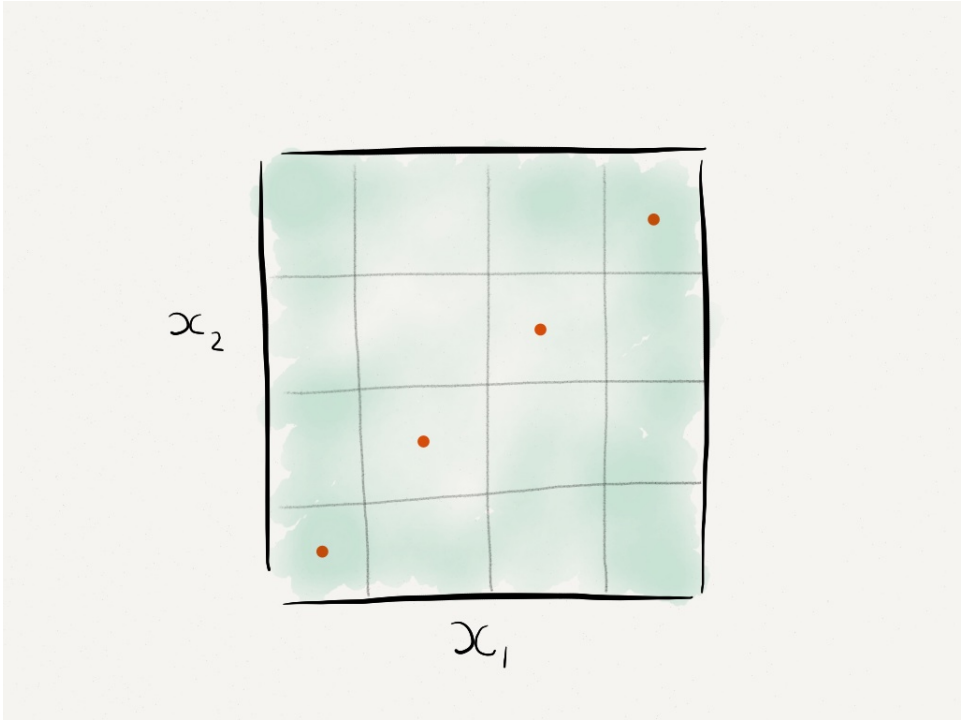
Divide input space into equally-probable regions ...



Maximin Latin Hypercube

An aside on experiment design

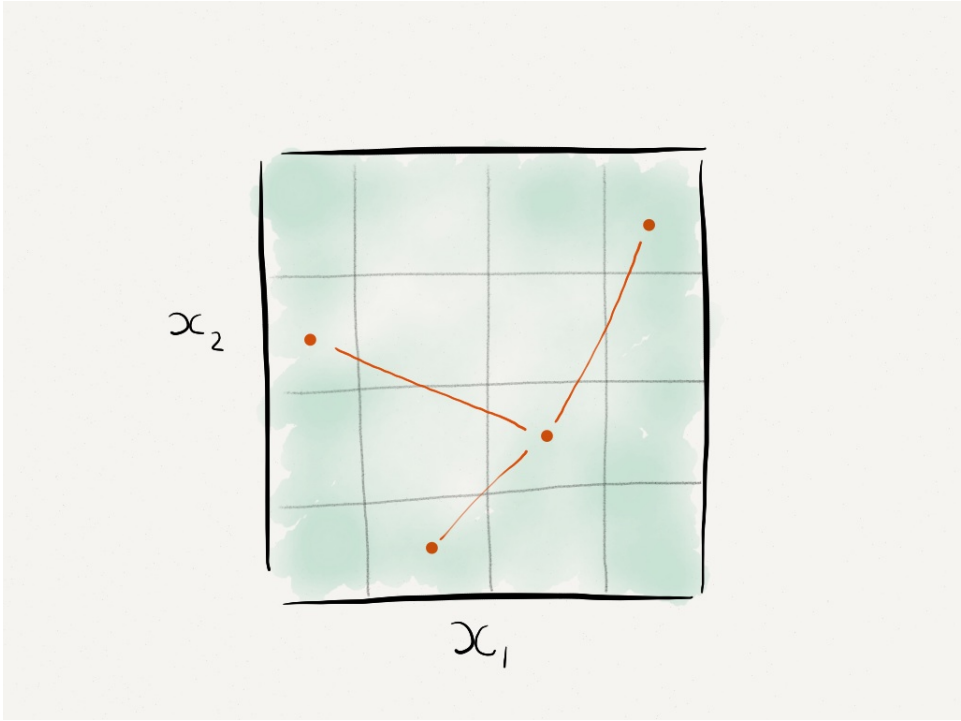
Make sure that each row and column has a point. Not like this!



Maximin Latin Hypercube

An aside on experiment design

Find the design with the largest minimum distance.



Unsampled uncertainties

We are limited when sampling uncertain quantities by:

- time: finite PhD and post-doc years, project end dates.
- computing power: finite resources.

But also by:

- imagination: finite awareness (of the limitations of our models and of ourselves).

Parameters, structures and inputs we didn't know were important. Values we didn't realise were physically meaningful.

This *will* happen.

Unsampled uncertainties (cont.)

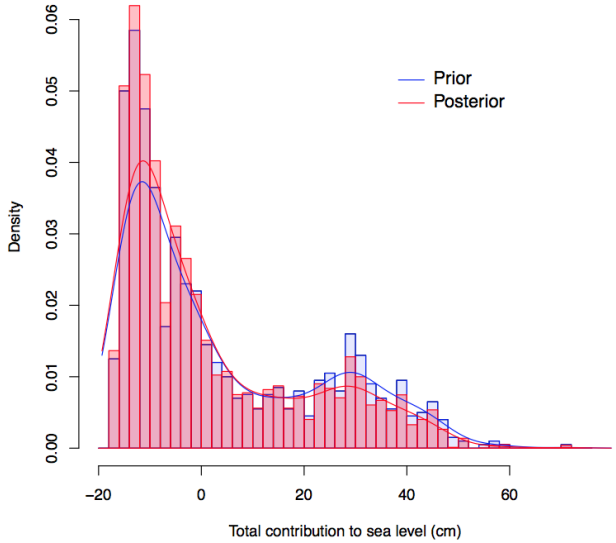
Try to plan for the effects of unsampled uncertainties:

- Do a screening study varying more parameters, with wider ranges, than you first planned
- Save computing resources to do simulations you wished you'd done
- Where possible, try structures used in other simulators
- Incorporate an estimate of structural uncertainty; remember it will be an underestimate.
- Be humble about the limitations of your uncertainty assessment.

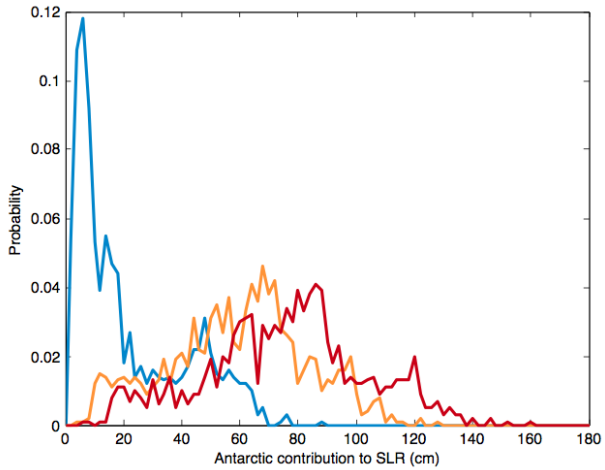
Your uncertainty assessments will always be inadequate. Try to make them the least inadequate you can. Strive to make them less inadequate throughout your career. Never give up!

Expect surprises!

Example: Antarctic ice sheet model ensemble.



More time $\rightarrow$ two more plausible structures:



Section B: Observational Comparison

Outline:

1. Choose outputs
2. Construct a distance measure
3. Deal with correlation
4. Incorporate structural uncertainty
5. (Set implausibility threshold)
6. Visualise the results

Choose outputs

Identify the 'crucial' outputs that are best at discriminating good from bad candidate values for θ^*.

If you are aiming to make probabilistic statements, choose variables for which simulator discrepancies are expected to be independent (often not the case for climate), otherwise you must attempt to quantify how they are correlated (hard).

In general it is easier to use a small number of outputs (fewer judgements; more likely to be independent). This might mean we choose to use summaries of observations.

Construct a distance measure

Define the distance between observations and simulator outputs at a given θ.

An observation z_i with observational error ϵ_i can be described by:

$$z_i = f_i(\theta^*) + \delta_i + \epsilon_i,$$

where $\mathrm{sim}(\theta^*)$ is now $f(\theta^*)$ and disc is δ.

We assume simulator discrepancy and observational error are uncorrelated, $\mathrm{Var}(\delta + \epsilon) = \mathrm{Var}(\delta) + \mathrm{Var}(\epsilon)$. So the natural, statistically-founded, distance measure for a given simulator output is:

$$d_i(\theta)^2 = \frac{(z_i - f_i(\theta))^2}{\mathrm{Var}(\delta_\mathrm{i}) + \mathrm{Var}(\epsilon_\mathrm{i})}$$

where $d_i(\theta)$ is the normalised Euclidean distance between two values.

Distance measure for multiple outputs

The main challenge with choosing a distance measure for multiple outputs is that simulator discrepancy is systematic in s_i (variable, location and time).

i.e. $\mathrm{system}(s_i; x) - f(s_i; x, \theta)$ is correlated with $\mathrm{system}(s_j; x) - f(s_j; x, \theta)$ when $s_i \neq s_j$.

Simply summing the squared discrepancies over *all* observations over-penalises candidate values for θ^*, because it double-counts the same discrepancy.

Formally, adding together squared residuals is *only a valid statistical distance measure* if the discrepancies are uncorrelated.

Dealing with correlation

If we are using history matching to rule out regions of parameter space without stating confidence intervals, i.e. we are not aiming to make probabilistic statements, we can use a simple method for combining distance measures even if they are correlated.

Quite popular: $d(\theta) = \max_i \{d_i(\theta)\}$.

Constructing independent tests

But if we:

- are aiming to make probabilistic statements, and
- wish to avoid statistically modelling the discrepancy correlations across outputs (and we almost always do),

we must construct a distance measure such that the discrepancies are probabilistically independent.

If you attempt to make your discrepancies independent, you are doing better than many (most?) other studies. Be wary when interpreting results from other studies that quantify simulator-observation differences without explicitly removing or statistically modelling the correlations.

We have already said that variables should be chosen so the discrepancies are independent. For independence in space and time, there are two options: clumping and thinning.

Clumping

Aggregate observations within large-scale regions and/or time intervals to create a smaller number of pseudo-observations (e.g. mean, trend, maximum).

Regions must be larger than the decorrelation length, and time intervals longer than the autocorrelation lag (e.g. continental and multi-annual scales).

Try to construct informative pseudo-observations.

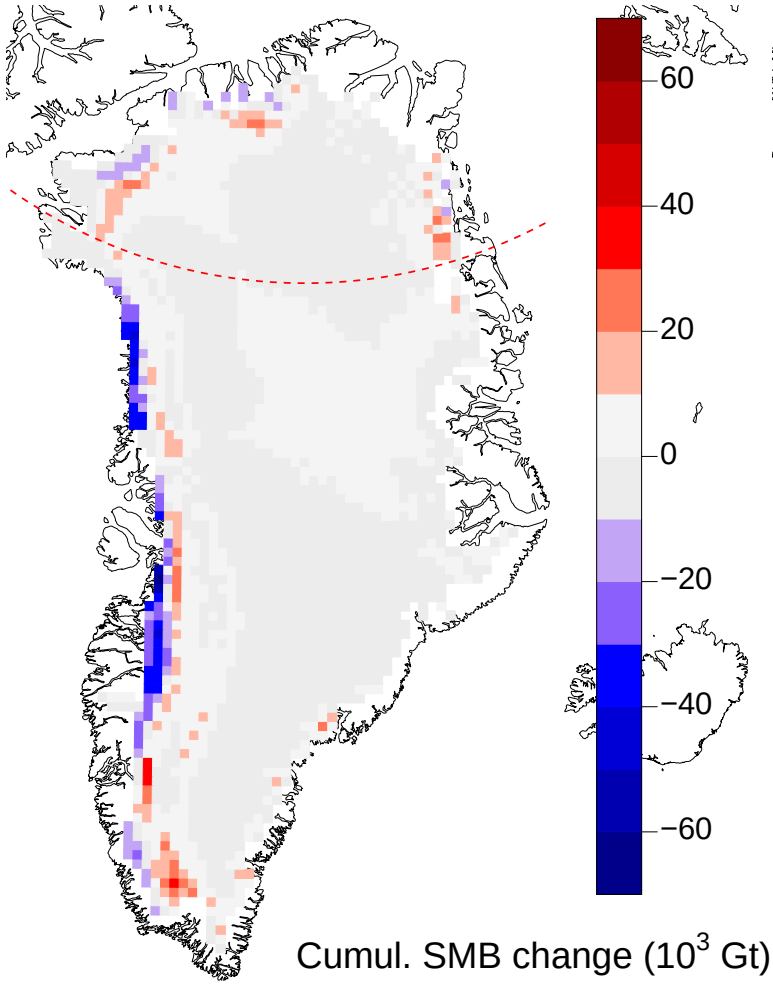
Thinning

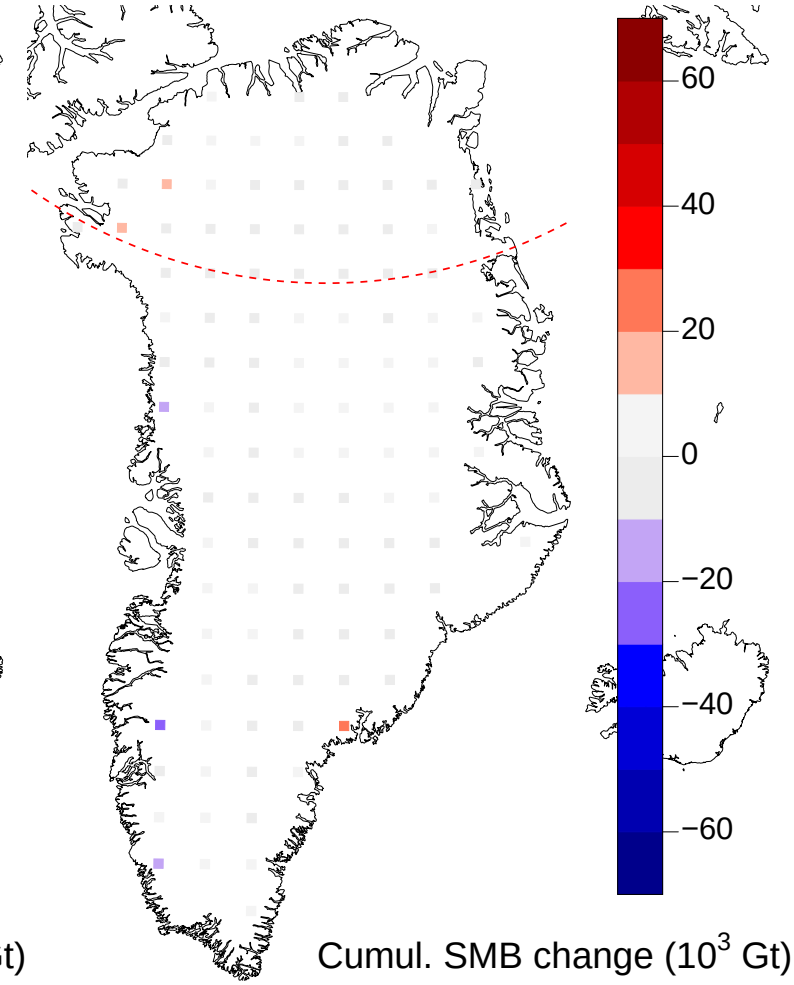
Use a subset of the observations.

The spacing between observations must be larger than the decorrelation length / autocorrelation lag.

Try to retain the most informative observations.

Example





What might have been a better choice?

Edwards et al. (2013)

Distance measure for independent discrepancies

For a given variable we now have observations $z = (z_1, z_2, ..., z_m)$, where $j = 1, ..., m$ indexes space and time for the clumped/thinned observations of that variable.

If we are using history matching to rule out regions of parameter space with confidence intervals, i.e. we are aiming to make probabilistic statements, we must combine our independent distance measures.

Unfortunately there is no simple rule, only ad-hoc ones.

One approach is to retain the individual $d_j(\theta)$ but adjust the implausibility threshold to account for multiple testing (see Inference section).

An alternative is to decide to use only one output after all...

Distance measure for independent discrepancies

If we are doing Bayesian calibration we can use

$$d(\theta)^2 = \sum_{j=1}^{m} \frac{(z_j - f_j(\theta))^2}{\mathrm{Var}(\delta_\mathrm{j}) + \mathrm{Var}(\epsilon_\mathrm{j})}$$

where $d(\theta)$ is the normalised Euclidean distance for multiple uncorrelated pairs of values.

If the observational errors have the same variance across space/time, replace $\mathrm{Var}(\epsilon_\mathrm{j})$ with $\mathrm{Var}(\epsilon)$.

Ditto $\mathrm{Var}(\delta_\mathrm{j})$.

Distance measure for correlated discrepancies

Advanced! If we wish to use a distance measure for correlated discrepancies, we can use the most general form: the Mahalanobis distance:

$$d(\theta) = (z - f(\theta))^T (\Sigma_\delta + \Sigma_\epsilon)^{-1} (z - f(\theta))$$

If you squint you can see this is still the normalised sum of squared discrepancies, but now the normalisation term is a matrix.

We need a matrix to represent discrepancy covariances in space/time:

$$\Sigma_\delta = \begin{pmatrix} \sigma^2_{1,1} & \sigma^2_{1,2} & \cdots & \cdots \\ \sigma^2_{2,1} & \sigma^2_{2,2} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \sigma^2_{m,m} \end{pmatrix}$$

Distance measure for indep. discrepancies, again

The normalised Euclidean distances are simplified versions of the Mahalanobis distance.

When the discrepancies are independent the covariance matrix is diagonal:

$$\Sigma_\delta = \begin{pmatrix} \sigma^2_{1,1} & 0 & \dots & \dots \\ 0 & \sigma^2_{2,2} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \sigma^2_{m,m} \end{pmatrix}$$

Observational variances are typically assumed to be diagonal too, so the distance measure collapses to the sum of squared discrepancies normalised by the variances for each location/time.

Distance measure for identically-distributed discrepancies, again

And if the discrepancy variances are independent and identically-distributed (constant variance in space/time), or 'i.i.d.', all the diagonal elements are equal so we can simplify more:

$$\Sigma_\delta = \sigma_\delta^2 \begin{pmatrix} 1 & 0 & \dots & \dots \\ 0 & 1 & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & 1 \end{pmatrix}$$

If the observational errors are also i.i.d., the distance measure collapses to the sum of squared differences normalised by the sum of the two variances.

Setting the discrepancy variance

Let's assume we are avoiding covariances.

We must set a discrepancy variance Var(δ) for each:

- variable
- region, if errors are likely to vary systematically in space
- time, if errors are likely to vary systematically in time

We may be able to approximate Var(δ) from comparisons of the model with observations.

Some studies estimate Var(δ) from comparisons of the MME with each other and/or observations.

Or we can set it with expert judgement and treat it as an unconstrained parameter. This may seem unnerving, but it is not as unnerving as the alternative...

The importance of the discrepancy variance

If we do not include a discrepancy variance, by ignoring it or setting it to zero, we are implicitly making the assumption that *we could tune a model to perfectly simulate the real world*.

This is ludicrously indefensible.

Any value is better than zero. If you put a non-zero number for the discrepancy variance, you are doing much better than many (the majority?) of the studies in the literature.

Be wary when interpreting studies in which you do not see an estimate of simulator discrepancy, e.g. a distance measure $d_j(\theta)^2 = \frac{(z_j - f_j(\theta))^2}{\mathrm{Var}(\epsilon_j)}$, because...

All models are wrong, even at their best parameter values. We *must* strive to quantify how wrong.

Discrepancy variance from observations

We can approximate variance(s) for independent discrepancies from the differences between the simulator and observations.

Pukelsheim 3 sigma rule

Call the total error variance $\sigma^2_{tot} = \sigma^2_\delta + \sigma^2_\epsilon$.

At least 95% of the discrepancies for the simulation at its best parameter values, $(z_j - f(\theta^*))$, will be within $\pm 3\sigma_{tot}$, i.e. $d_j(\theta^*) < 3$ for most j. This is true for any continuous unimodal probability distribution (Pukelsheim, 1994).

So we can choose σ_δ such that this is true.

We might use a single value everywhere (σ_δ), or different values for different regions/times (σ^j_δ) if the discrepancies look systematically different.

Discrepancy variance from observations

Unfortunately we do not know θ^*!

Our best ensemble member, f($\tilde{\theta}$), will likely be a worse match to observations.

Setting σ_δ^2 such that at least 95% of discrepancies $d_j(\tilde{\theta}) < 3$ is likely to over-estimate σ_δ^2.

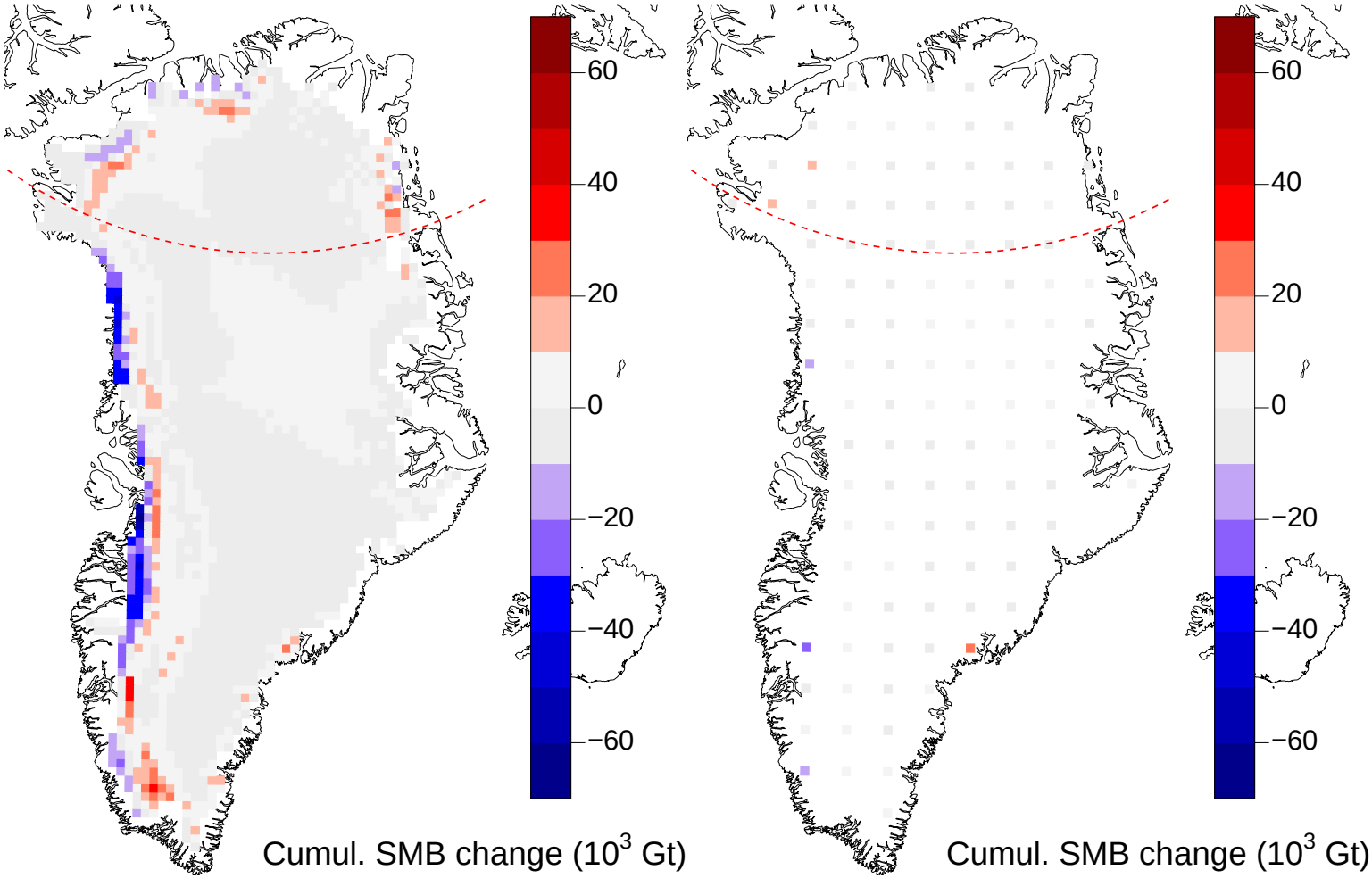
But if

- the parameter space is not too large
- the ensemble size is reasonable
- the simulator is not highly variable across the parameter space

then f($\tilde{\theta}$) provides a reasonable first order starting point.

We also need to have sufficient number of observations to be reasonably confident in our estimate.

Fixed discrepancy variance



I used $\sigma_\delta = 20x10^3$ Gt.

What might have been a better choice?

Discrepancy variance from MME

Two recent approaches to estimating a full discrepancy covariance matrix for HadCM3:

- UKCP09. Sexton et al. (2011) use the differences between HadCM3 and the MME members
- RGH* model. Williamson et al. (2013) use the differences between HadCM3 and the MME mean and the differences between the MME mean and observations

*Rougier, Goldstein and House.

To specify a full covariance matrix, you probably need to find a friendly** statistician.

**They mostly are.

Discrepancy variance from expert judgement

If we cannot estimate Var(δ) from observations *or* an MME because we have:

- too few ensemble members, too large a parameter space, a highly variable simulator
- too few observations
- too few MME members (e.g. for non-climate models) or nearby statisticians

then we can set discrepancy variances using expert judgement alone.

Discrepancy variance from expert judgement

Questions we might consider for this:

- How well do developers and users expect the simulator to perform?
- How large are the observational uncertainties? (simulator discrepancy must be larger!)
- How many ensemble members are ruled out (HM) or down-weighted (BC) using the value chosen? (see Inference section)

This is an unconstrained parameter. So it is sensible to check how much your results are affected by different values.

Implausibility threshold

In history matching we set a threshold c to rule out regions of parameter space as implausible; $d(\theta) > c$ indicates that θ is implausible as a candidate for θ^*.

A popular choice is $c = 3$ using Pukelsheim (McNeall et al., 2013; Williamson et al., 2013); $d(\theta^*) < 3$ with a probability greater than 0.95;

Ensemble members that pass the threshold are judged not implausible or not ruled out yet (NROY).

Recap

Section A: (Initial) Experimental Design

- Choose uncertain quantities to sample
- Express these as probability distributions
- Sample efficiently

Section B: Observational Comparison

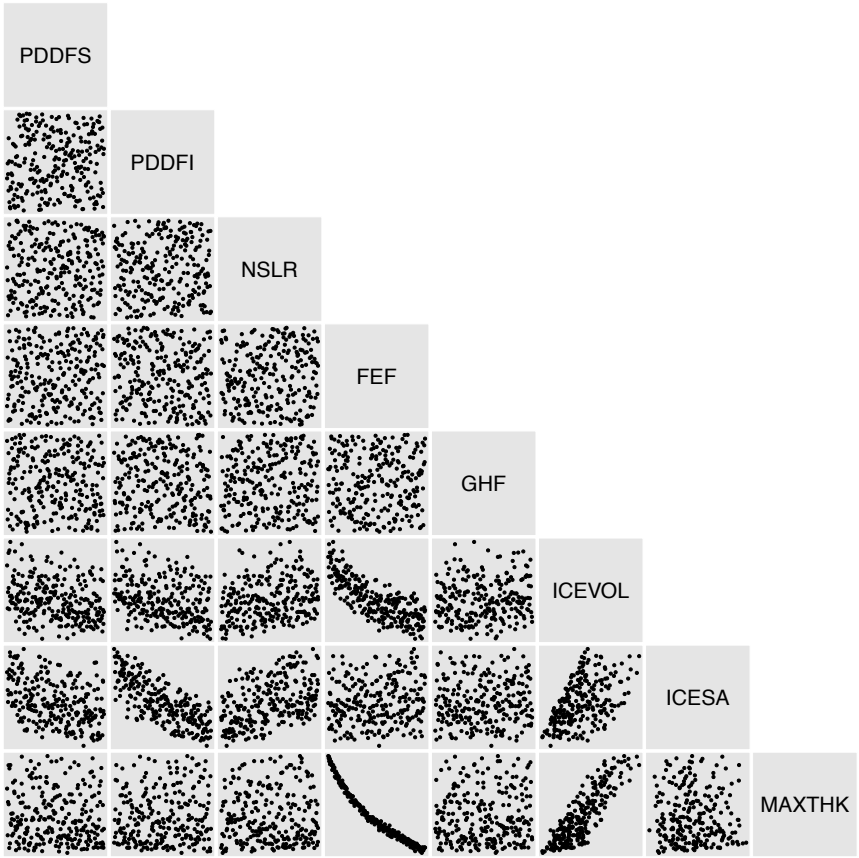
- Choose outputs
- Construct a distance measure
- Deal with correlation
- Incorporate structural uncertainty
- (Set implausibility threshold)
- Visualise the results

Visualisation: pairs plot

A popular way to visualise relationships is with a pairs plot.

In the following diagram the first five rows are parameters of a Greenland ice sheet model, and the last three are summary outputs (McNeall et al., 2013).

Exploratory analysis pairs plot



Visualisation: pairs plot

Useful:

- show the observations on the output axes
- plot the distance measure and implausibility threshold

But remember: high-dimensional spaces are not intuitive to humans, e.g. 1-D and 2-D visualisations may suppress almost all the information, or even mislead.

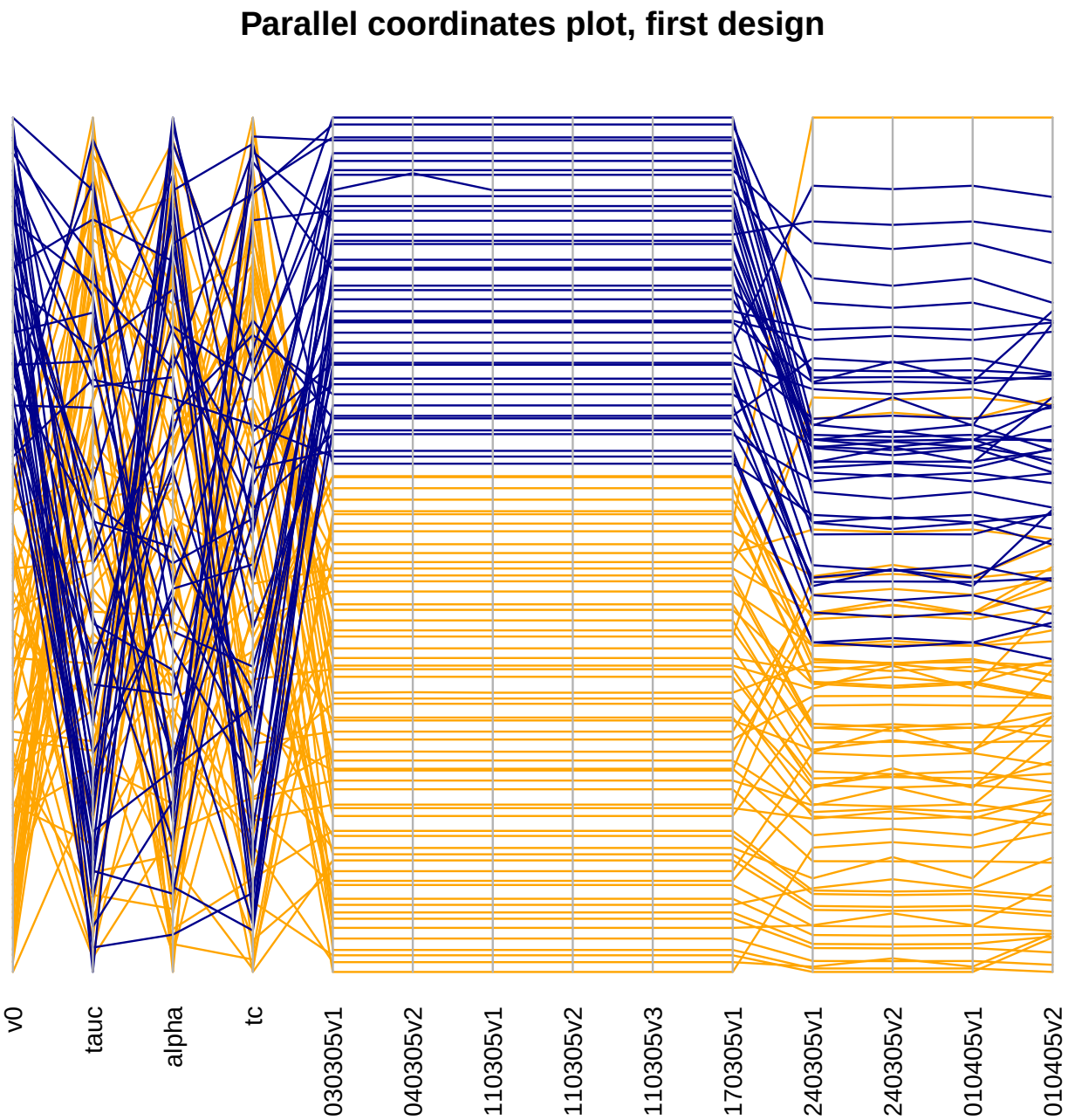
	$\mathbf{X=0}$	$\mathbf{X=1}$	$\mathbf{X=2}$	
$\mathbf{Y=0}$	0.20	0.10	0.00	0.30
$\mathbf{Y=1}$	0.05	0.15	0.15	0.35
$\mathbf{Y=2}$	0.05	0.15	0.15	0.35
	0.30	0.40	0.30	1.00

Visualisation: parallel coordinate plot

A very useful tool for high-dimensional data is the parallel coordinate plot.

In the following diagram the first four rows are parameters of an avalanche model, and the last ten are outputs.

Ensemble NROY members are shown in dark blue. At least one parameter has a large fraction of its range ruled out.


Parallel coordinates plot, first design
v0
tauc
alpha
tc
030305v1
040305v2
110305v1
110305v2
110305v3
170305v1
240305v1
240305v2
010405v1
010405v2

Visualisation: parallel coordinate plot

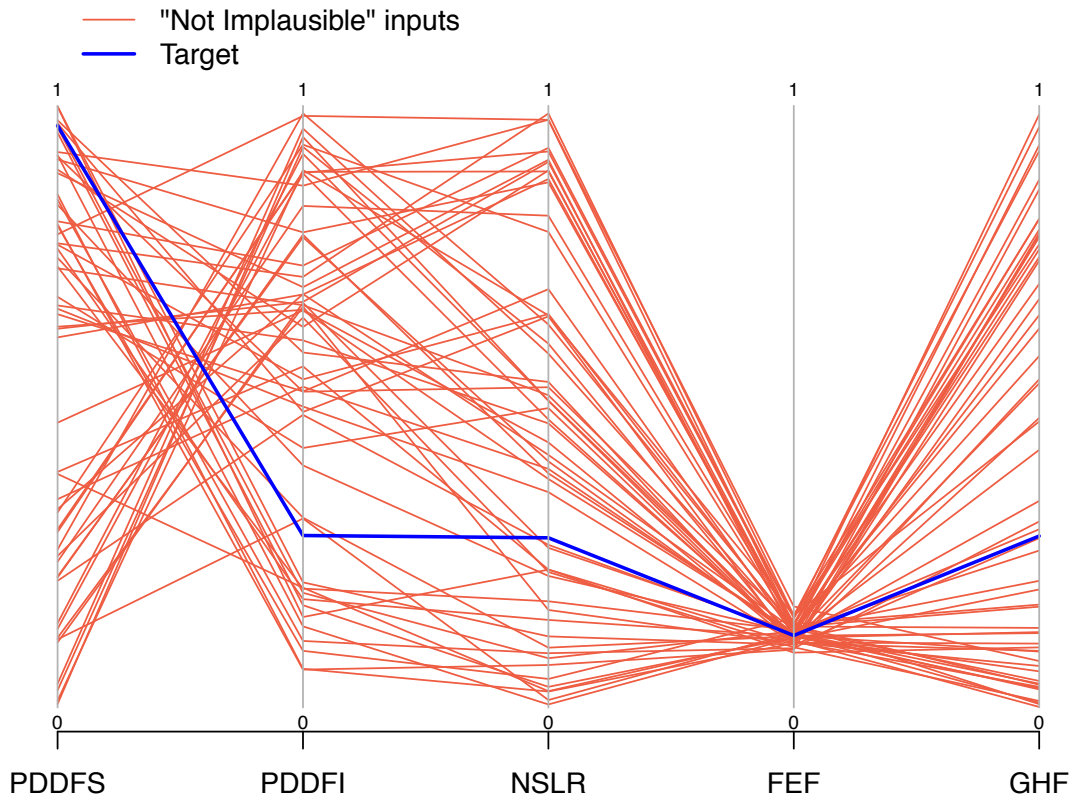
Look for clusterings, relationships, human errors and compensating parameters.

- Plot/highlight NROY ensemble members
- Try flipping one or more variables to look for inverse relationships
- Try re-ordering the variables to untangle the spaghetti

The following diagram shows a PCP for the parameters of the NROY members of the Greenland ensemble, along with the parameter values of the 'observations' (in this case another simulation, as a test).

Most of the range of one parameter is ruled out. Two parameters seem to compensate each other!

Parallell coordinates plot



Increase the ensemble size

If possible use a sequential experimental design: use batches of simulations to successively refine the ranges of the candidate values for θ^*.

This will help use your computing resources on the good/important parts of parameter space.

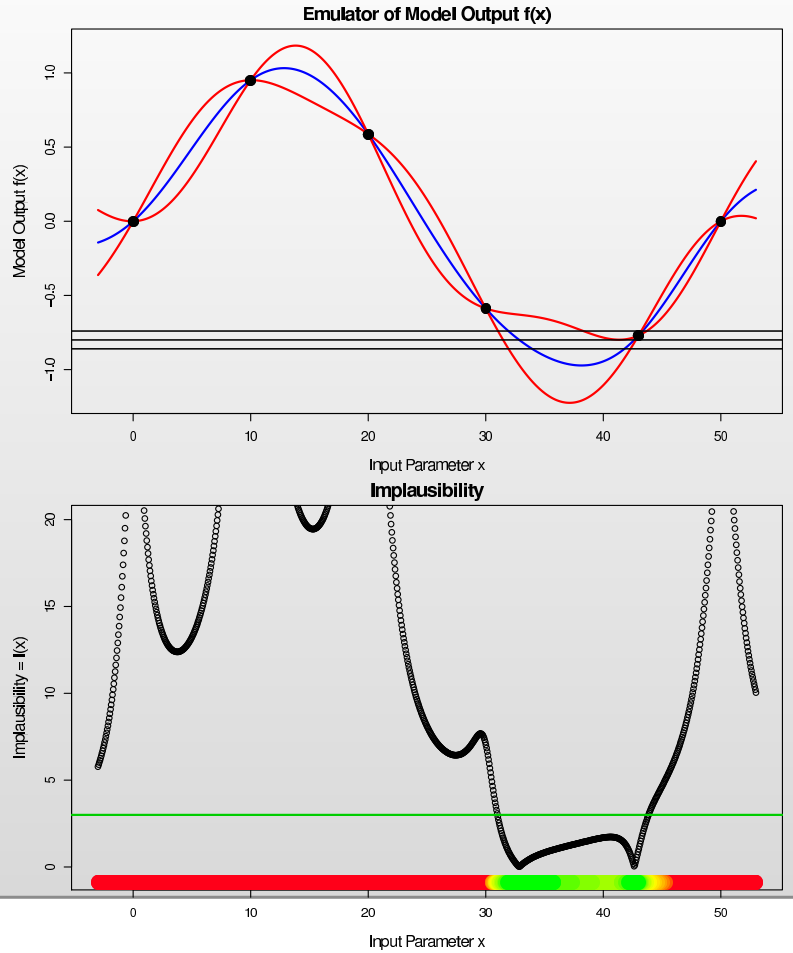
Visualise the ensemble to test which parameters and outputs are doing most of the work.

Refine the ranges and do a second ensemble. Visualise again.

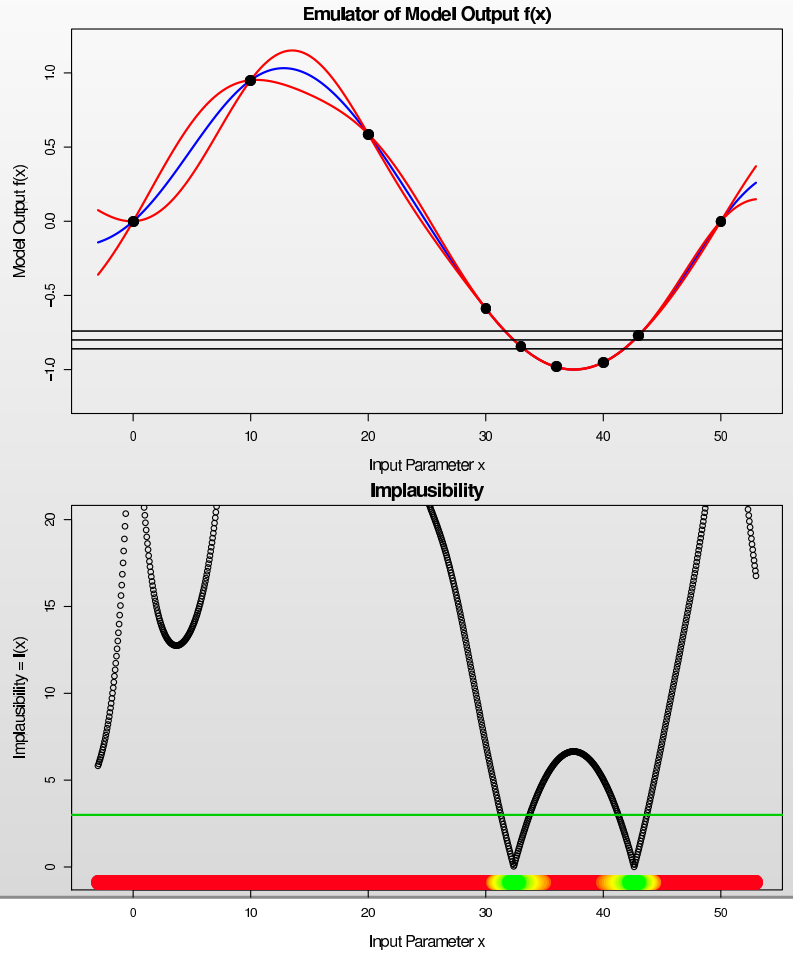
Repeat until the budget of runs is exhausted (or you are exhausted).

The following slides demonstrate how effective a sequential design can be (Vernon et al., 2010).

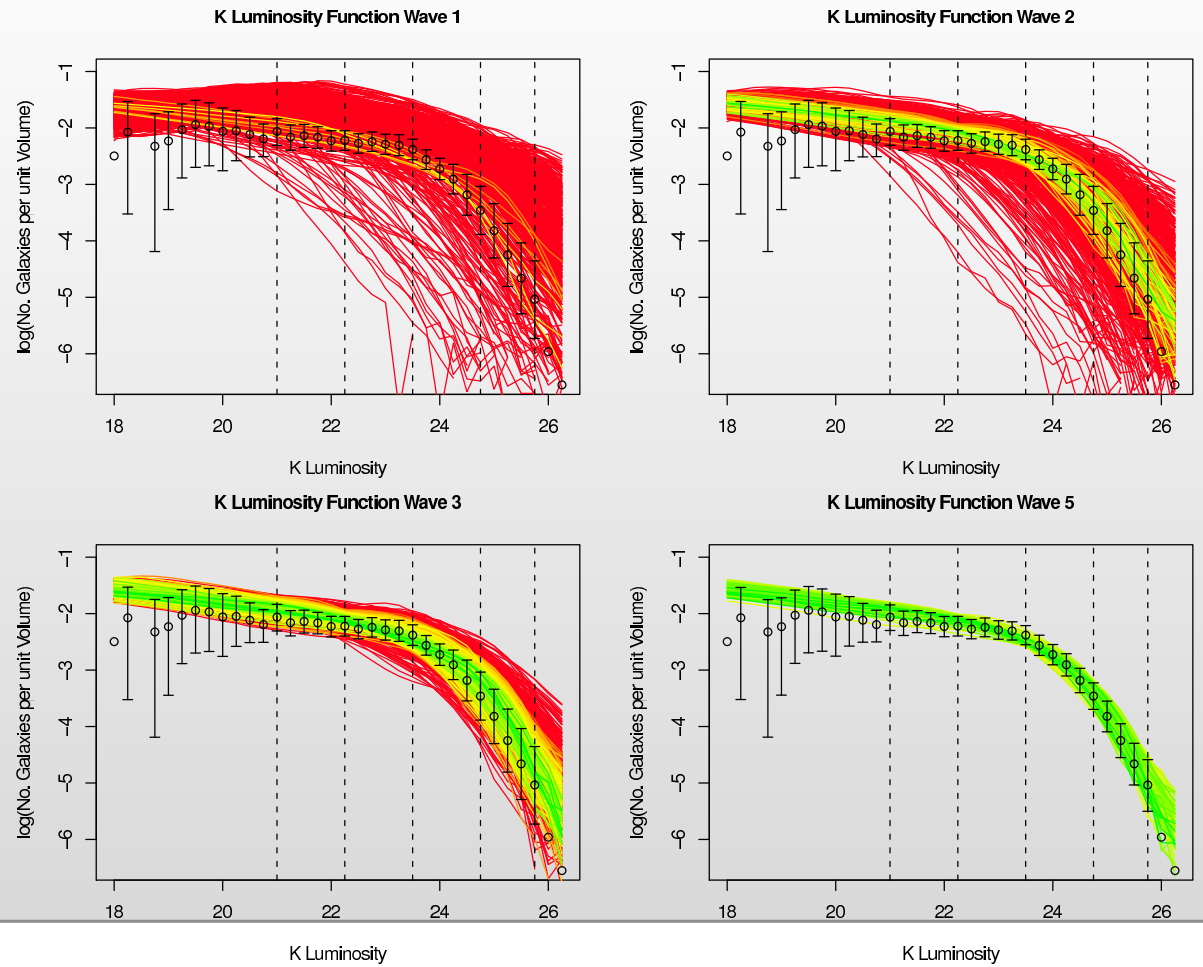
History Matching via Implausibility: a 1D Example



History Matching via Implausibility: a 1D Example



bj Luminosity Output of Waves 1,2,3 and 5



Refining the ranges

Look at the relationships between inputs and outputs in the pairs plot and parallel coordinates plot.

- Are there any strong relationships between parameters and outputs?

If one parameter is doing all the work, consider zooming in for that one and possibly zooming out for the others.

If none of the parameters are doing much, consider expanding the range of them all!

Refining the ranges

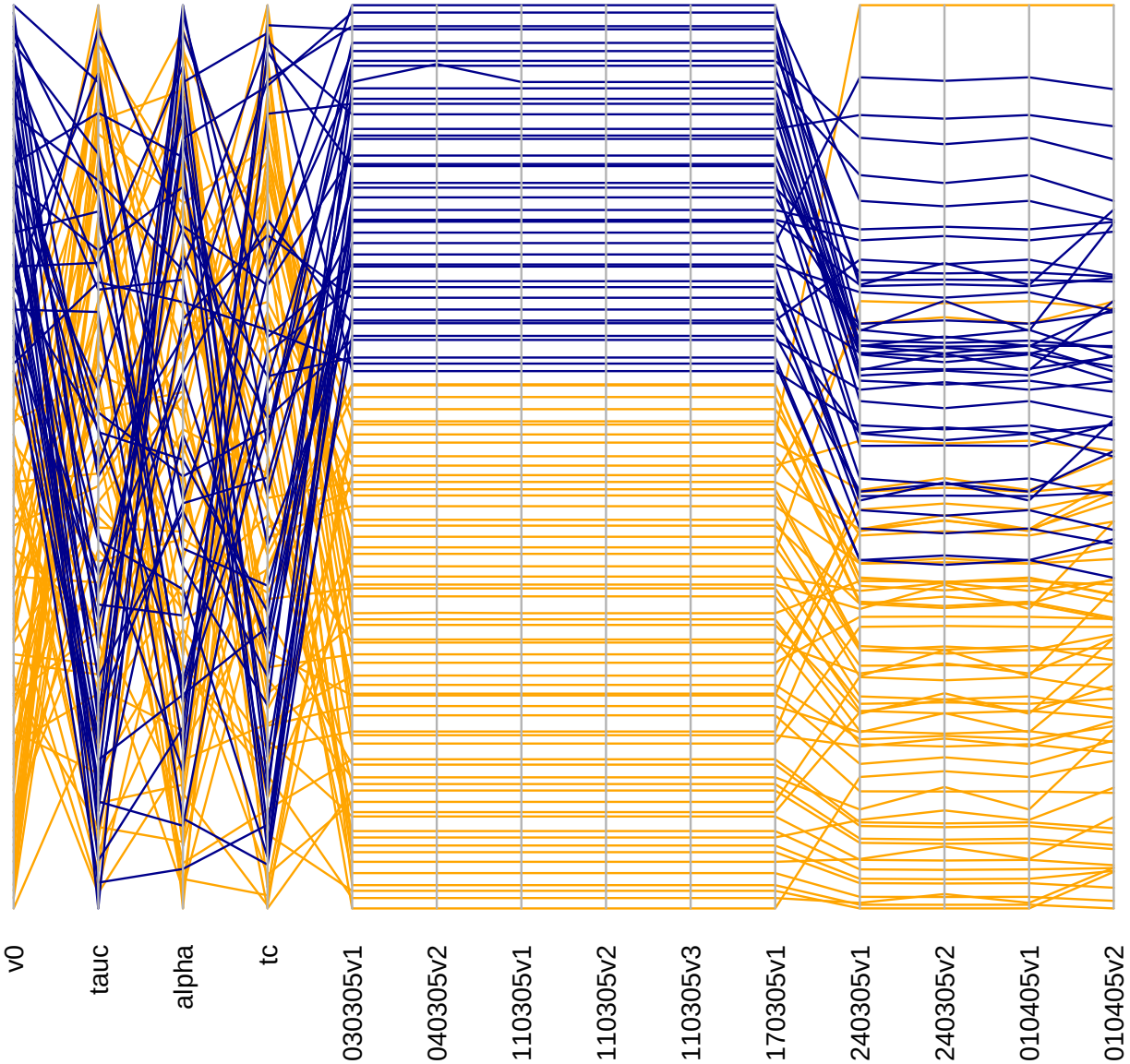
Look at the relationship between outputs and observations:

- Plot/highlight NROY ensemble members, or colour by values of the distance measure.
- Are there any NROY simulations?
- Do they pile up at one end of one or more of the parameters?

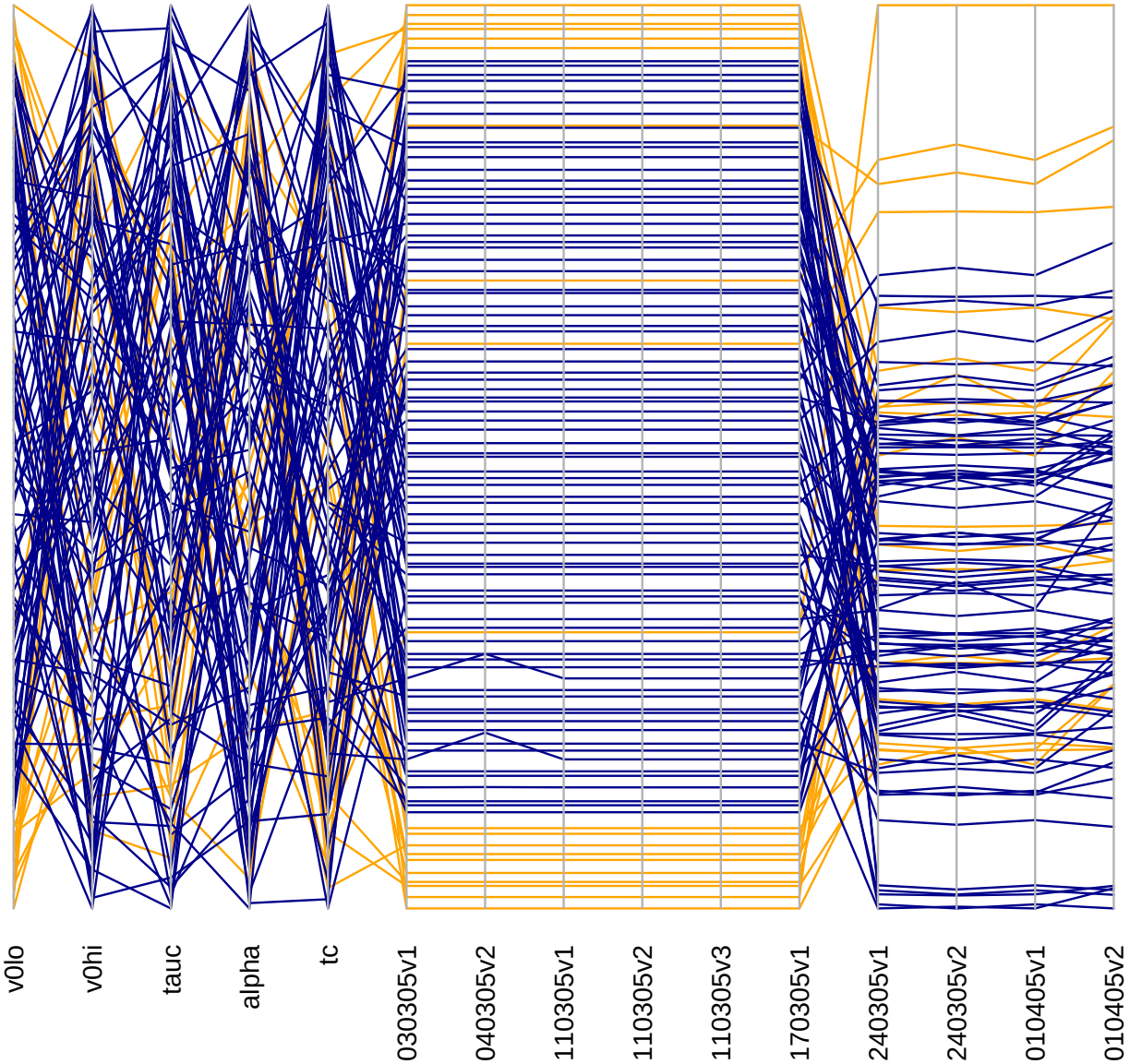
If the NROY values are at one end of a parameter range, expand the range on one side.

If the NROY values are in the middle of a parameter range, consider restricting the range, especially if some values are producing non-physical outputs.

Parallel coordinates plot, first design



Parallel coordinates plot, second design



Sequential analysis

With each new batch of runs, revisit:

- the output of the best ensemble member $f(\tilde{\theta})$
- the discrepancy variance, if estimated from $f(\tilde{\theta})$, and therefore the distance measure (this may change $\tilde{\theta}$)
- the subset of the ensemble not ruled out yet

Sampling sequentially

It is hard to *extend* a LHC to a larger ensemble size: it is much easier to do a new, separate LHC.

In theory this is less efficient, but in a high-dimensional space we probably won't get points close together.

An alternative would be a Sobol sequence, which is. . . sequential, therefore easy to extend. But it is much harder to code, and therefore more vulnerable to error.

Section C: Statistical Inference

Overview:

- History matching with confidence intervals
- Bayesian calibration

History matching

Obtaining CIs using implausibility for a single output is easy! Applying $c = 3$ gives the candidate values for θ^* not ruled out according to a 95% CI.

But an overall CI from m multiple outputs is not: we must deal with the effect of multiple testing. There is no universally agreed way to do this.

Gladstone et al. (2012) construct a joint test with significance level 5% from m independent tests of significance level α, where α is calculated with $1 - 0.05 = (1 - \alpha)^m$. From the new, smaller α they calculate a new, larger c (Pukelsheim, 1994) which they apply to all distance measures d_j.

N.B. The Pukelsheim rule is conservative (e.g. 99.7% of normal distribution lies within 3 s.d. of the mean); it only rules out very poor candidates for θ^*.

Bayesian calibration

If you:

(a) have constructed your distance measure(s) as described;
- Euclidean distance for independent discrepancies
- Mahalanobis distance otherwise (difficult)

(b) have a reasonable ensemble size;

(c) are comfortable assuming your discrepancies are normally distributed (not necessarily!);

then it is easy to do a Bayesian update.

(c) is an assumption of a Gaussian likelihood function...

Bayesian calibration

. . . luckily, your distance measure is the exponent of a Gaussian.

So for each ensemble member with parameter values θ, and for each variable, you can calculate a weight:

$$w(\theta) = \exp\{-\tfrac{1}{2}\mathrm{d}(\theta)\},$$

where we discard the constant at the front of the Gaussian because we now normalise the weights across the N ensemble members:

$$w'(\theta) = \frac{w(\theta)}{\sum_N w(\theta)}$$

If you chose variables such that the discrepancies are independent, you can simply multiply the weights for each variable together.

If not, you need to (a) use only one variable, or (b) find a statistician.

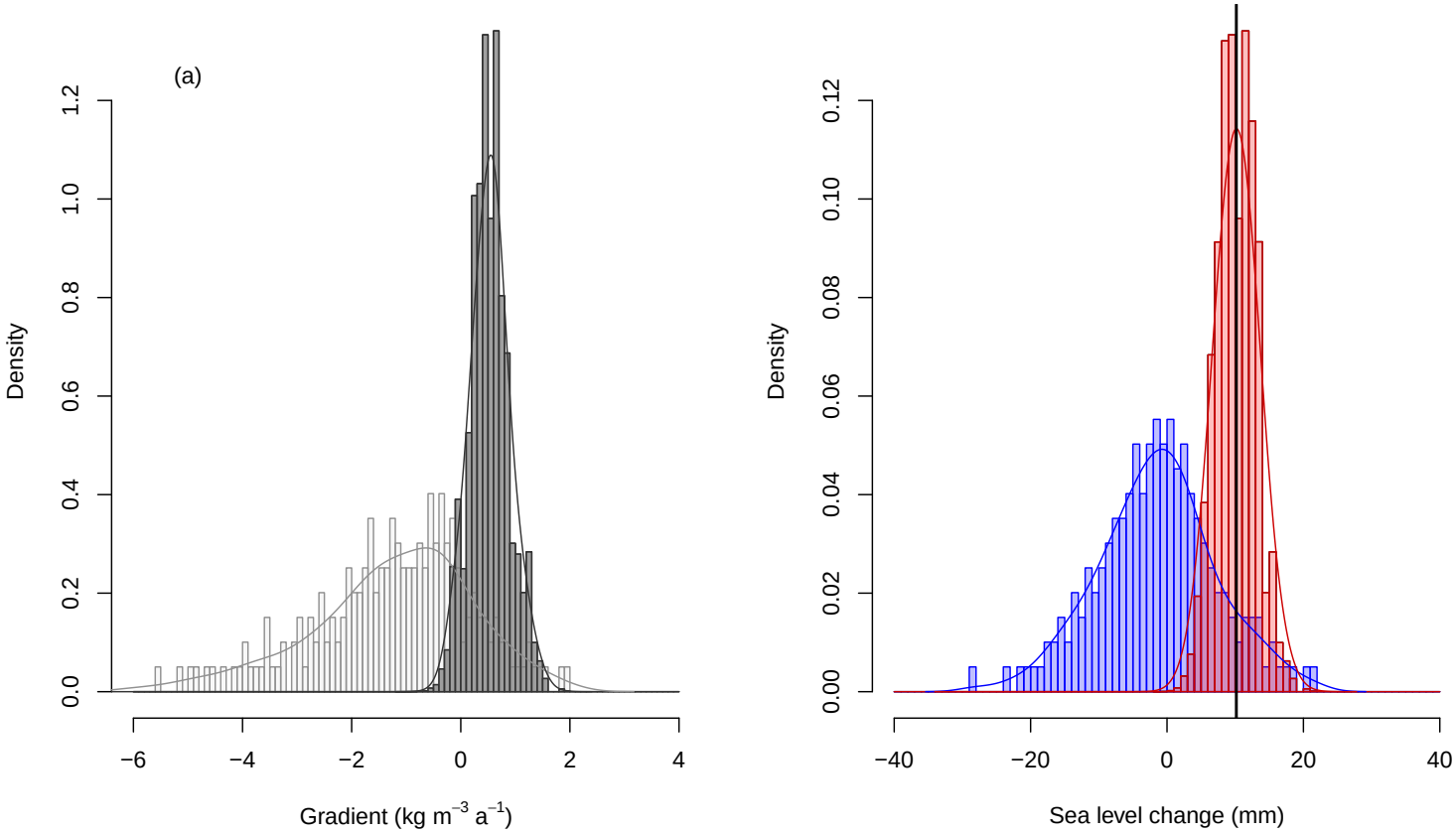
Bayesian calibration

Inspect the weights to see if they are all concentrated on one member, or very evenly spread; alter the discrepancy variance to test the effect.

If our ensemble size is not too small, we can estimate:

- posterior distributions by plotting weighted histograms of the parameter values and simulator outputs for which we want to quantify uncertainty (e.g. projections);
- credibility intervals by resampling from the ensemble (bootstrapping) with the weights to estimate sample quantiles;
- probability densities, for example the maximum probability value, using kernel density estimation of the histograms with the weights.

Example


(a)
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
−6
−4
−2
0
2
4
Gradient (kg m^{-3} a^{-1})
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
−40
−20
0
20
40
Sea level change (mm)

Emulation

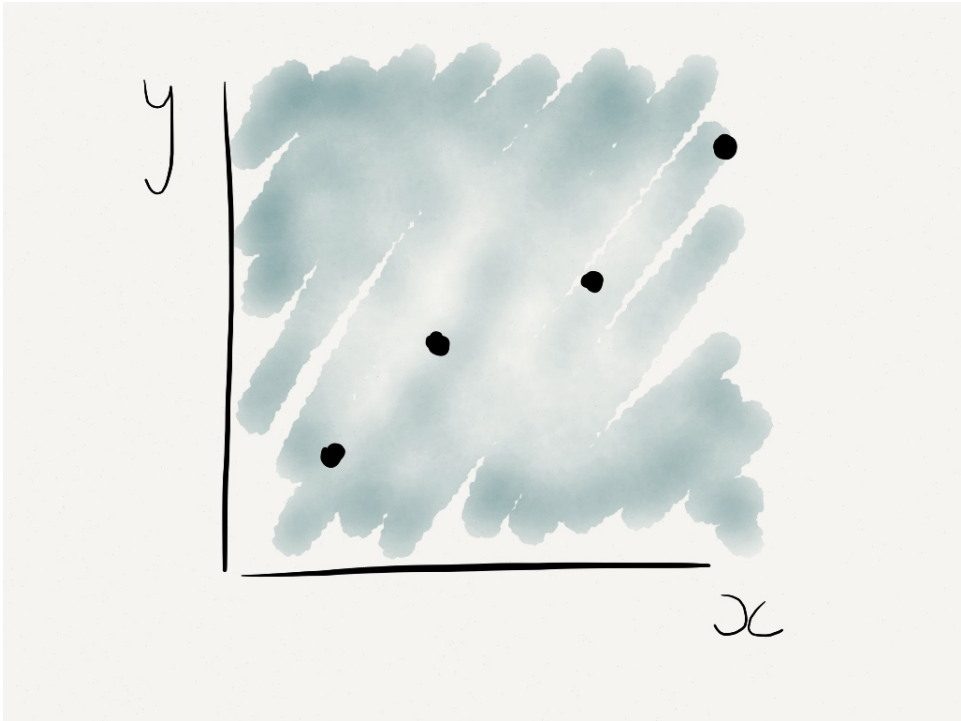
My simulator's too expensive! to sample multiple parameters or estimate a probability distribution.

Emulation:

- Statistical modelling of the output of a simulator as a function of its parameters
- Mere mortals are limited to univariate emulation (one output at a time).
- A 'Gaussian process' emulator is a natural place to start for deterministic, smooth-ish simulators.
- R packages: e.g. BACCO, DiceKriging, SAVE
- MUCM Toolkit and UCM email list: `http://www.mucm.ac.uk`

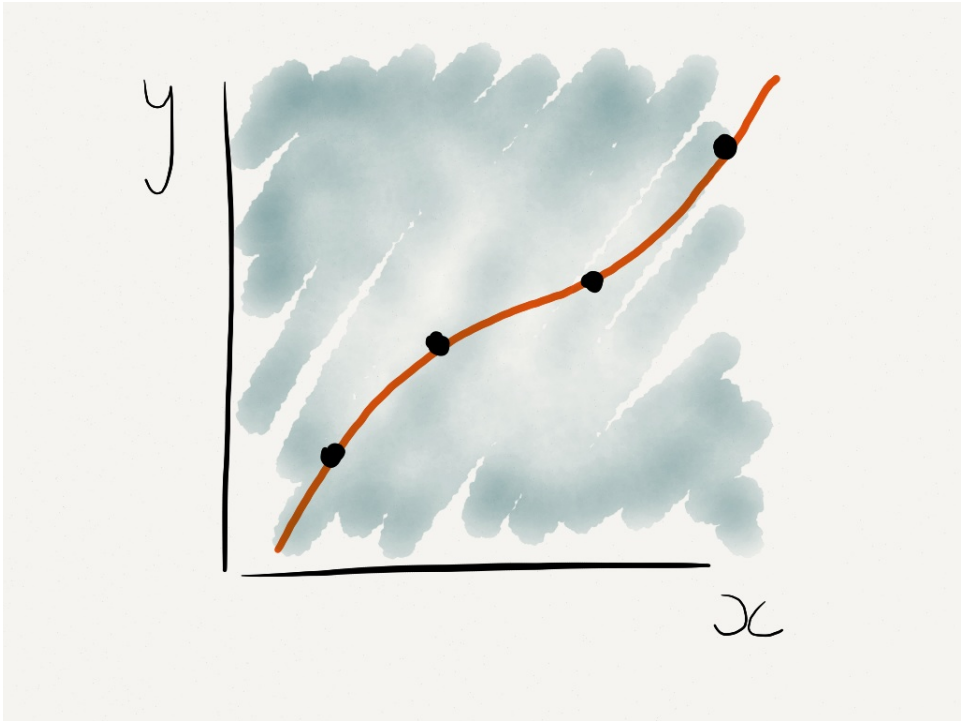
Building an emulator

Run a small set of design points.



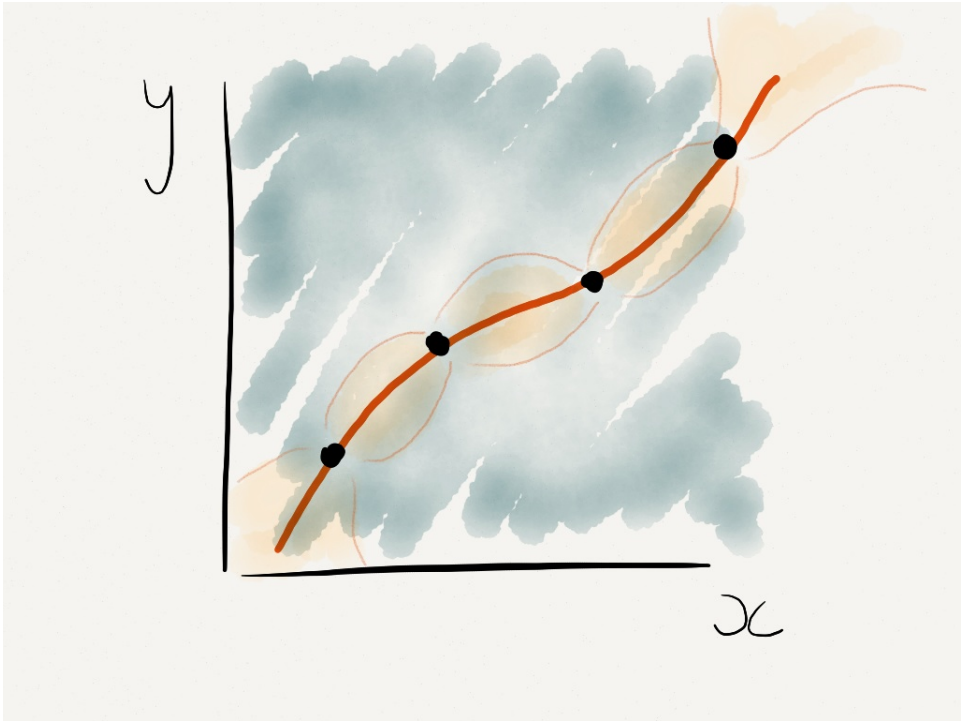
Building an emulator

The emulator fits a smooth mean function through the points ...



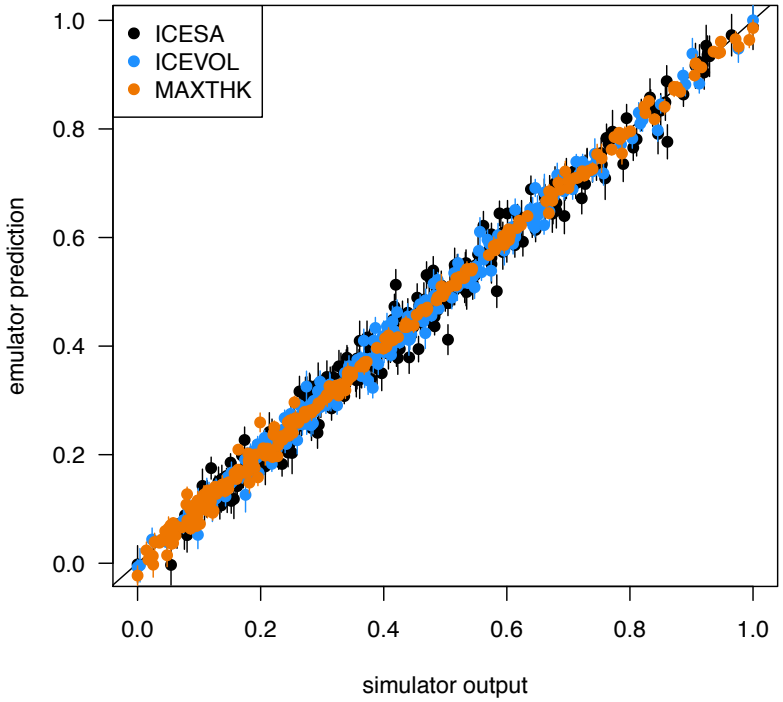
Building an emulator

... and includes an estimate of uncertainty.



Checking the emulator

- Use a leave-one-out prediction diagnostic to check the emulator is working.
- Perfect predictions lie on the straight line.



Summary

- History matching:
  1. rule out implausible parameter values
  2. (optional:) estimate confidence intervals for not implausible parameter values
- Bayesian calibration:
  3. estimate probability distributions for parameters

Here be dragons:

- Specifying prior probability distributions (3)
- Dealing with correlated discrepancies across variables, space, time (2, 3)
- Setting implausibility threshold (1, 2)
- Joint inference across multiple outputs (2, 3)

Summary

Here be dragons (all):

- Statistically-justifiable distance measure / likelihood function
- Setting discrepancy variance
- Sequential ensemble design

Rewards:

- Reduced uncertainties (all)
- Quantified uncertainties (2, 3)

that are statistically meaningful and understandable.

Further reading

History matching:

- Antarctic glacier simulator:: Gladstone et al. (2012), Calibrated prediction of Pine Island Glacier retreat during the 21st and 22nd centuries with a coupled flowline model. Earth and Planetary Science Letters, vol 333-334., pp. 191 - 199
- Emulated galaxy formation simulator: Vernon et al. (2010), Galaxy Formation: a Bayesian Uncertainty Analysis. Bayesian Analysis 5:4, 619-670.
- Emulated Greenland ice sheet simulator: McNeall et al. (2013), The potential of an observational data set for calibration of a computationally expensive computer model, Geoscientific Model Development 6, 2369-2401.
- Emulated GCM: Williamson et al. (2013), History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. Climate Dynamics, doi: 10.1007/s00382-013-1896-4.

Further reading

Bayesian calibration:

- Emulated GCM: Sexton et al. (2012), Multivariate probabilistic projections using imperfect climate models part I: outline of methodology. Climate Dynamics 38:11-12, 2513-2542; Sexton and Murphy (2012), Multivariate probabilistic projections using imperfect climate models. Part II: robustness of methodological choices and consequences for climate sensitivity. Climate Dynamics 38:11-12, 2543-2558.
- Greenland ice sheet parameterisation: Edwards et al. (2013), Effect of uncertainty in surface mass balanceelevation feedback on projections of the future sea level contribution of the Greenland ice sheet Part 1: Parameterisation. The Cryosphere Discussions, 7, 635-674, 2013.

Further reading

MME mean discrepancy:

- RGH model: Rougier, J.C., M. Goldstein, and L. House (2013). Second-order exchangeability analysis for multi-model ensembles. J Am Stat Assoc, in press. `http://www.maths.bris.ac.uk/~MAZJCR/mme4.pdf`

Other:

- 3-sigma rule: Pukelsheim (1994). The 3-sigma-rule. Am. Stat. 48: 2, 88-91.
- Starter text: Rice (2007), Mathematical Statistics and Data Analysis, Brooks/Cole, Cengage Learning.
- Experimental design: Santer et al. (2003), The Design and Analysis of Computer Experiments, Springer Verlag.