Multivariate emulation for North American mid-Holocene temperature reconstructions

Jonathan Rougier

Department of Mathematics
University of Bristol, UK

with Tamsin Edwards, Mat Collins,
and other members of the PalaeoQUMP team

ECMWF, June 2011

Climate simulators and their parameters

$$\text{HadCM3}: \left.\begin{array}{l}\text{Parameters}\\ \text{Boundary conditions}\end{array}\right\} \mapsto \text{climate simulation.}$$

- We would like to
  - Understand better the behaviour of HadCM3,
  - Propagate parametric uncertainty into projections, and
  - Tune the parameters, which are under-determined.

Climate simulators and their parameters

$$\text{HadCM3} : \left.\begin{array}{l}\text{Parameters}\\ \text{Boundary conditions}\end{array}\right\} \mapsto \text{climate simulation.}$$

- We would like to
  - Understand better the behaviour of HadCM3,
  - Propagate parametric uncertainty into projections, and
  - Tune the parameters, which are under-determined.
- The challenges that we face:
  - A very long run-time (weeks/months per run)
  - Substantial internal variability in the outputs
  - Somewhat *ad hoc* collection of simulator runs.

Climate simulators and their parameters

$$\text{HadCM3} : \left.\begin{array}{l}\text{Parameters}\\ \text{Boundary conditions}\end{array}\right\} \mapsto \text{climate simulation.}$$

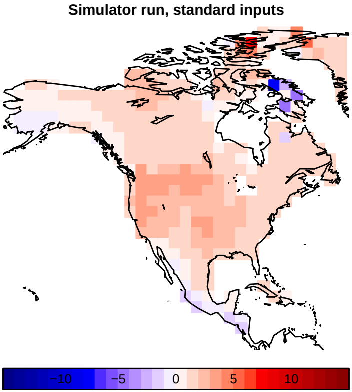
- We would like to
  - Understand better the behaviour of HadCM3,
  - Propagate parametric uncertainty into projections, and
  - Tune the parameters, which are under-determined.
- The challenges that we face:
  - A very long run-time (weeks/months per run)
  - Substantial internal variability in the outputs
  - Somewhat *ad hoc* collection of simulator runs.
- Statistical task: estimating the smooth function m in

$$\text{HadCM3}(r, \text{MH}) - \text{HadCM3}(r, \text{PI}) \equiv m(r) + \text{internal variability}(r)$$

where r is the parameter values, and PI and MH are Pre-Industrial and Mid-Holocene boundary conditions.

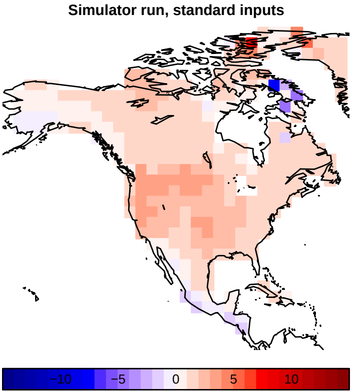
An ensemble of HadCM3 runs

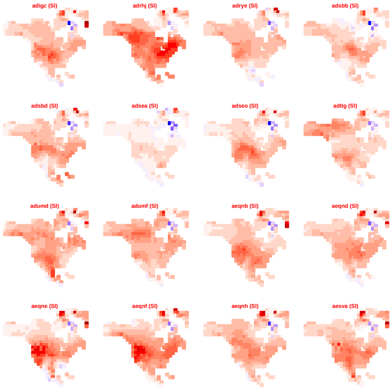
Each picture shows Had{C,A}M3's mid-Holocene North American MTWA anomaly for one setting of the simulator parameters.



An ensemble of HadCM3 runs

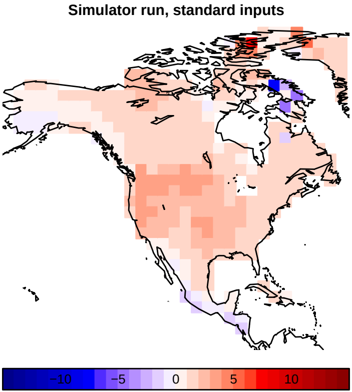
Each picture shows Had{C,A}M3's mid-Holocene North American MTWA anomaly for one setting of the simulator parameters.

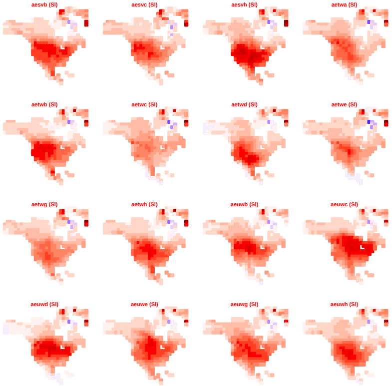




An ensemble of HadCM3 runs

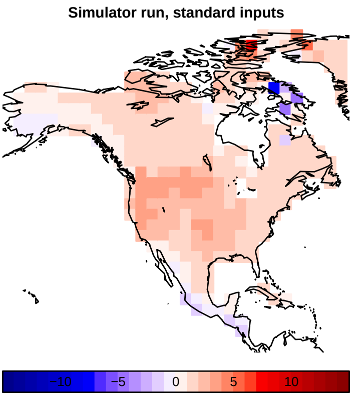
Each picture shows Had{C,A}M3's mid-Holocene North American MTWA anomaly for one setting of the simulator parameters.

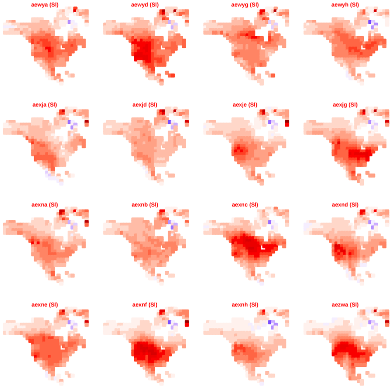




An ensemble of HadCM3 runs

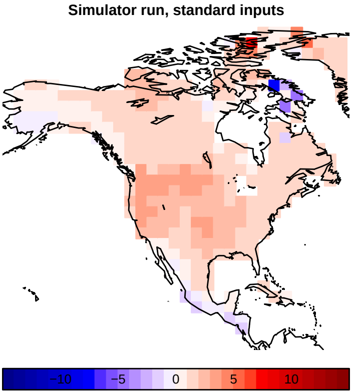
Each picture shows Had{C,A}M3's mid-Holocene North American MTWA anomaly for one setting of the simulator parameters.

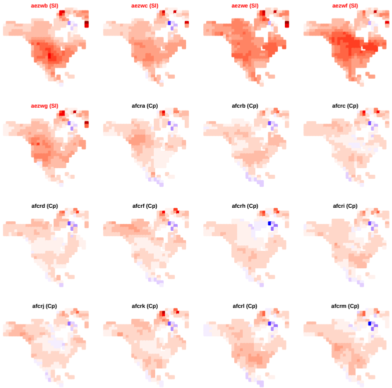




An ensemble of HadCM3 runs

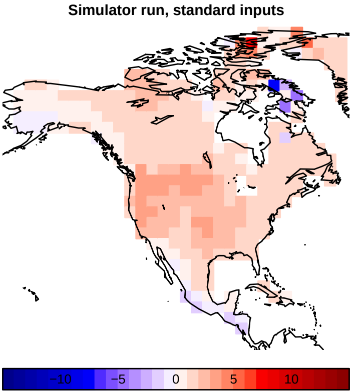
Each picture shows Had{C,A}M3's mid-Holocene North American MTWA anomaly for one setting of the simulator parameters.

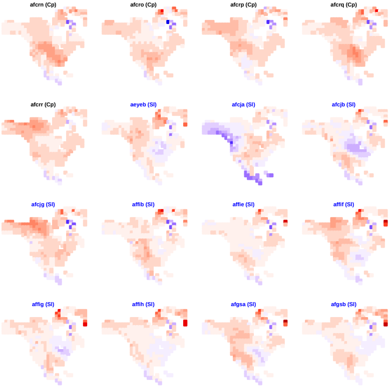




An ensemble of HadCM3 runs

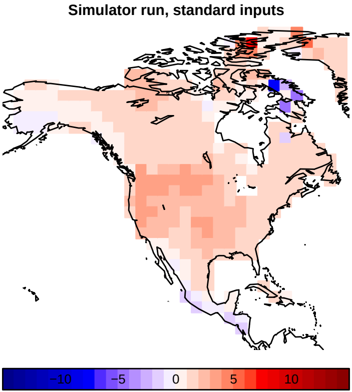
Each picture shows Had{C,A}M3's mid-Holocene North American MTWA anomaly for one setting of the simulator parameters.

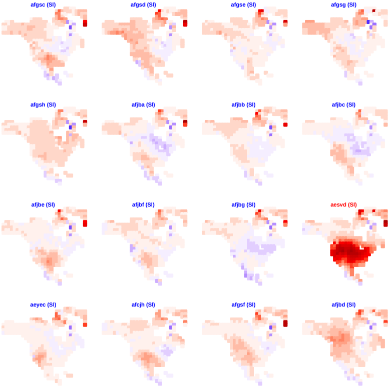




An ensemble of HadCM3 runs

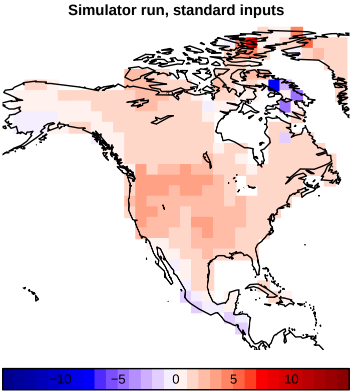
Each picture shows Had{C,A}M3's mid-Holocene North American MTWA anomaly for one setting of the simulator parameters.

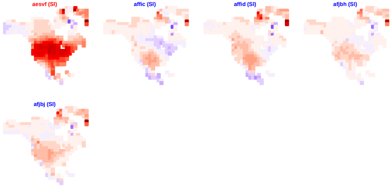




An ensemble of HadCM3 runs

Each picture shows Had{C,A}M3's mid-Holocene North American MTWA anomaly for one setting of the simulator parameters.





Five steps to an emulator for HadCM3

1. Consider the simulator $f(r)$ to be the sum of a smooth function plus internal variability, and estimate $S \approx \mathrm{Var}(\text{internal variability})$.

2. Dimensionally-reduce the smooth function, keeping only those linear combinations that we trust.

3. Estimate the mean and variance functions for the dimensionally-reduced smooth function, using the ensemble and S.

4. Recover the mean and variance of the smooth function in the full-dimension simulator output space.

5. Extensive full-dimension predictive diagnostic checking.

2. Dimensionally reduce the simulator output

2. Dimensionally reduce the simulator output

- Project the smooth component $m(r)$ onto the column-space of a matrix of basis vectors D (few columns), such that

$$\text{actual climate} \approx \bar{f}(r) := (DD^{+})^{T} m(r)$$

where D^{+} is the Moore-Penrose inverse of D.

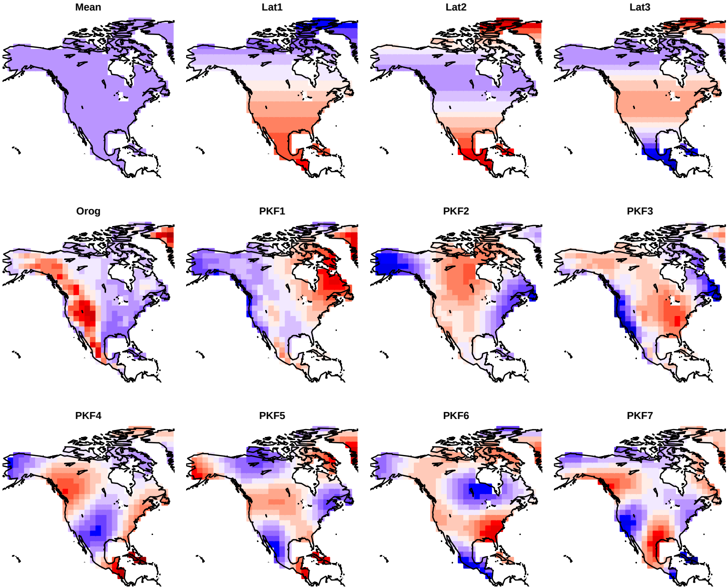
- We can, equivalently, write

$$\bar{f}(r) = (D^{+})^{T} D^{T} m(r) \equiv (D^{+})^{T} v(r) \qquad (\dagger)$$

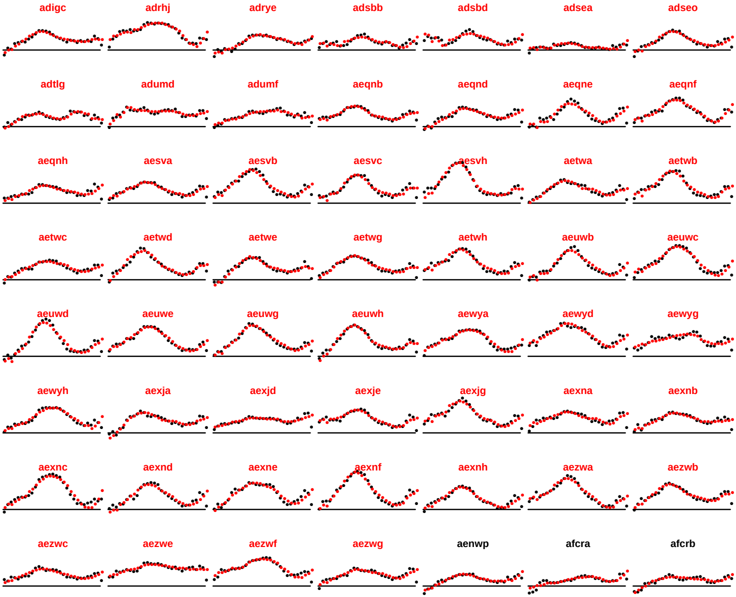
where $v(r) := D^{T} m(r)$ is a low-dimensional smooth function, effectively the coefficients of the linear combinations in $(D^{+})^{T}$.

- We will construct a mean function and variance function for v which we can then map into a mean function and variance function for $\bar{f}$, using $(\dagger)$.

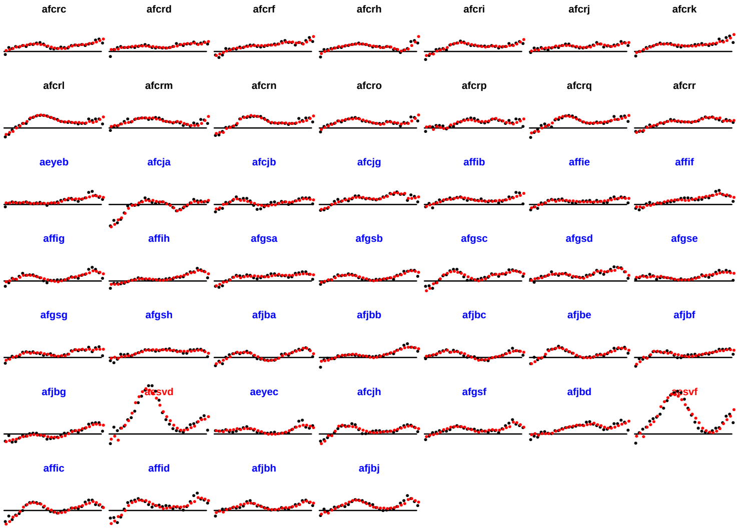
Our choice of filtering matrix, D



Our choice of filtering matrix, D



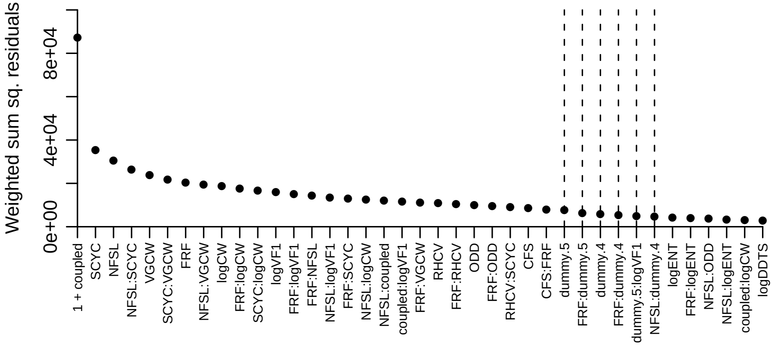
Our choice of filtering matrix, D



Choosing the regressors of $v(r)$

Choosing the regressors of $v(r)$

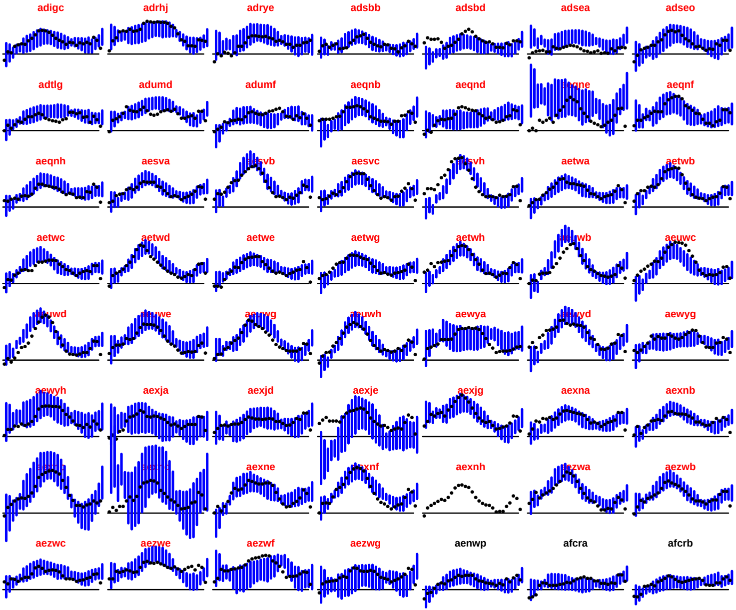
A simple approach here is to reduce the sum of squared residuals over the transformed variables, having adjusted for the covariance structure of the internal variability. Dummy regressors are used to check for over-fitting.



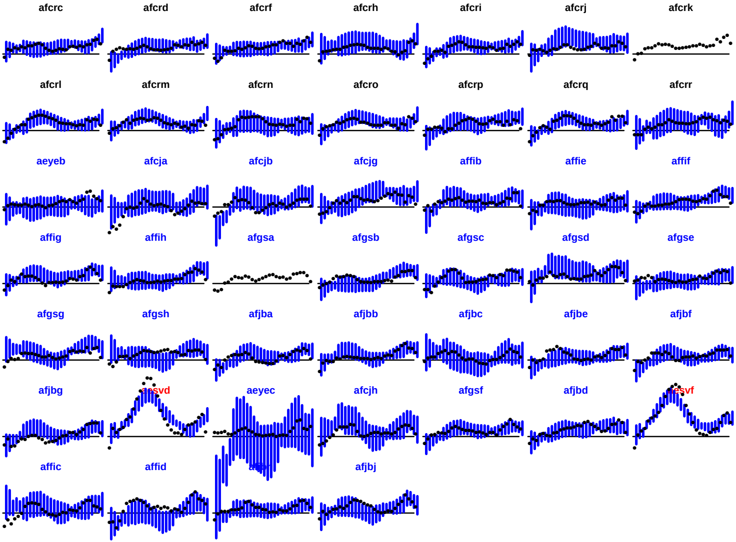
(Nb: we can get to 0 if we want to!)

Checking the emulator (LOO)

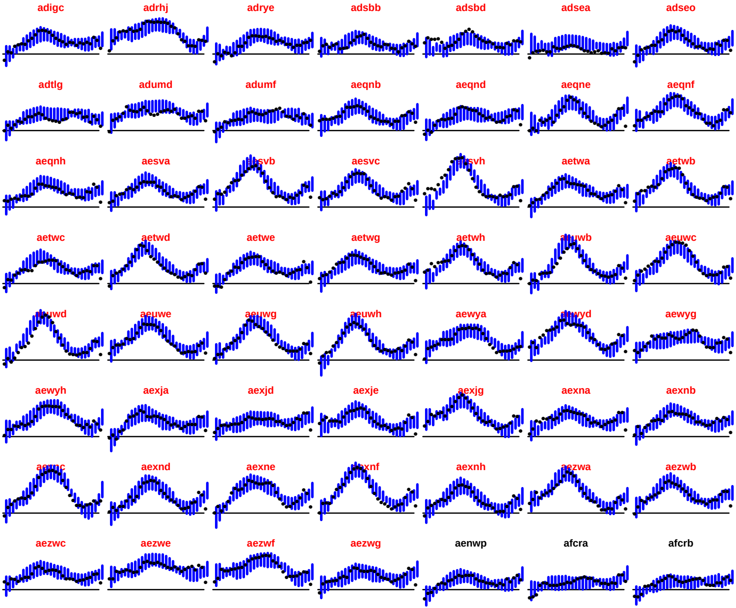
Checking the emulator (LOO)



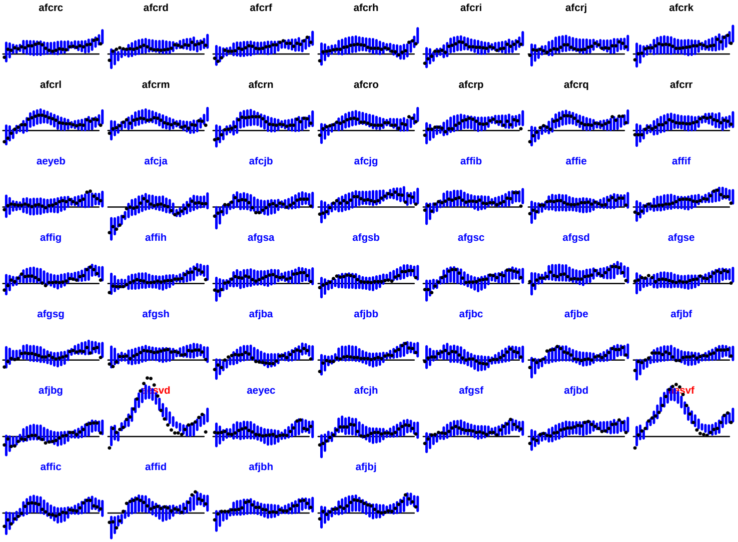
Checking the emulator (LOO)



Checking the emulator (LNO)



Checking the emulator (LNO)



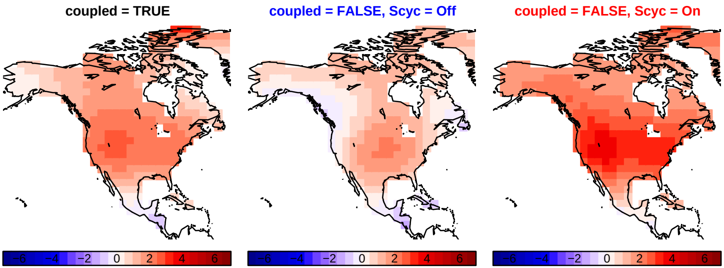
Dirty linen plot

'Collateral perturbations' in the experiment meant that switching on the sulphur cycle with a slab ocean removed an anthropogenic cooling that should not have been in a pre-industrial simulation.

Dirty linen plot

'Collateral perturbations' in the experiment meant that switching on the sulphur cycle with a slab ocean removed an anthropogenic cooling that should not have been in a pre-industrial simulation.

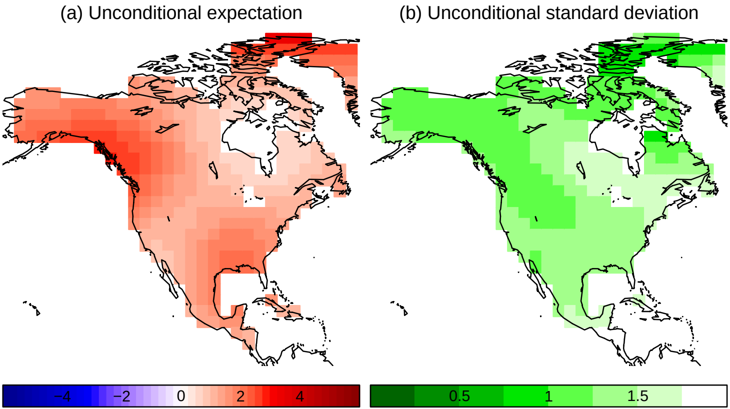
- This picture is *very tentative* and shows the emulator mean function with all other parameters at their standard settings.



Emulators are very important in identifying and adjusting for coding issues.

Integrating out the parameters

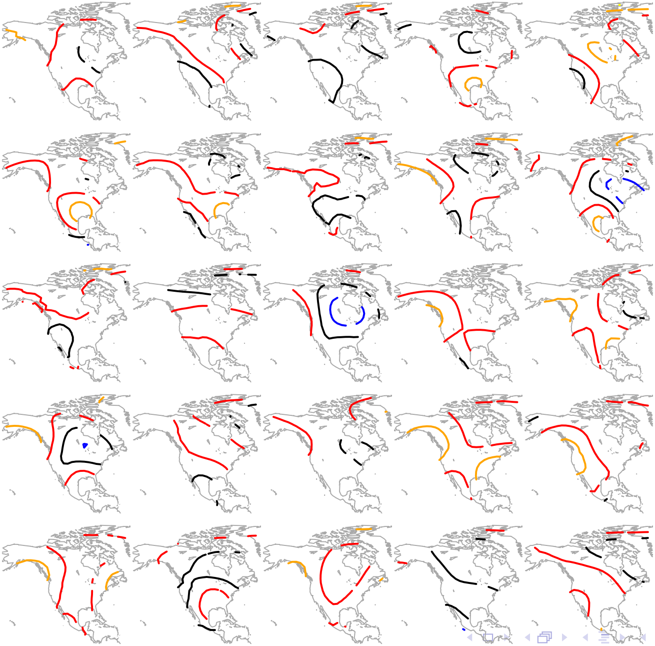
The emulator gives us a conditional mean and variance for the simulator output: *conditional* indicating that it depends on the choice of r. We can integrate the parameters out according to a specified distribution to find the *unconditional* mean and variance.



Integrating out the parameters

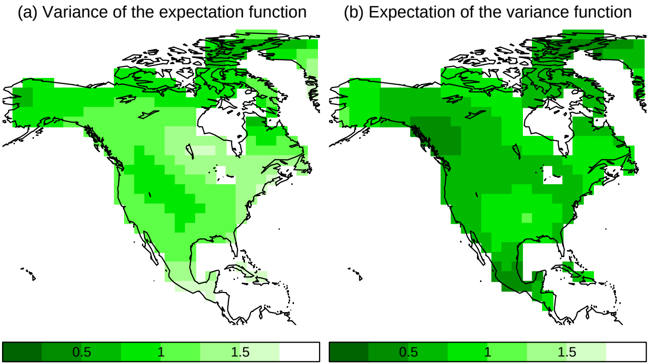
Sampling the isotherms gives a feeling for spatial variability. These isoterms are shown at -2°C (blue), 0°C (black), $+2^\circ$C (red), $+4^\circ$C (orange), and $+6^\circ$C (yellow).

Integrating out the parameters



Integrating out the parameters

There are two contributions to the unconditional variance: (a) the variability of simulator's response to the parameters, and (b) uncertainty about the simulator (the latter can be reduced with more runs). Here, it is the variability of the response which is contributing more.



Summary

REM: Statistics does not provide 'numbers'—*it provides a framework within which we can examine the impact of our judgements on our conclusions and actions*. One important role of this framework is to clarify the questions.

Climate simulator questions:

1. How to get a robust estimate of internal variability?
2. What linear combinations of high-dimensional spatial outputs are 'trustworthy'?
3. How to choose the regression functions for the simulator smooth component?

On the basis of our choices we compute an unconditional mean and variance for HadCM3 output, allowing for parametric uncertainty, and we attribute the variance primarily to parametric uncertainty, rather than not having enough simulator runs.

Acknowledgements

This is joint work with Tamsin Edwards at the University of Bristol, and Mat Collins at the University of Exeter and the Hadley Centre at the UK Met Office.

The support of other members of the QUMP and PalaeoQUMP projects is gratefully acknowledged, notably Philip Brohan, Michel Crucifix, Sandy Harrison (PI of PalaeoQUMP), and David Sexton.