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

MPI für Meteorologie, Hamburg, 16 Mar 2011

Palaeoclimate reconstruction

1. 'Pseudo-observations' based on proxy measurements have a high spatial resolution, but sparse coverage, and can be rather inaccurate

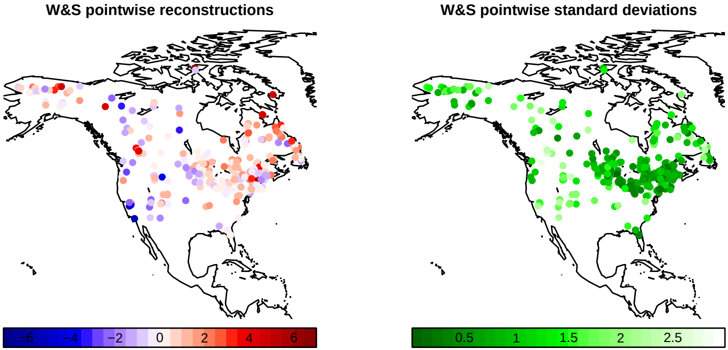
2. Climate simulator runs have full coverage but low spatial resolution, and there is the problem of simulator limitations

... Can we construct a synthesis of these two sources of information which combines their strengths?

This is a very generic problem. A *statistical* solution emphasises the assessment and role of uncertainty, represented probabilistically.

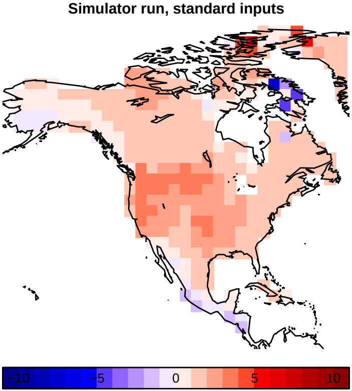
Pseudo-obs for pointwise reconstructions

Mid-Holocene MTWA anomalies.



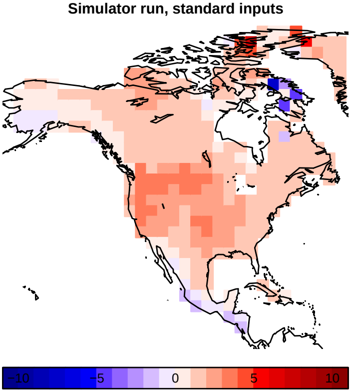
HadCM3 runs

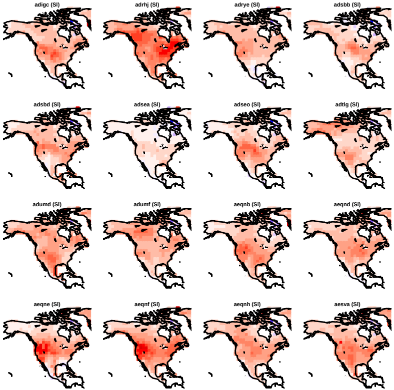
Standard parameterisation and some of our ensemble members (n.b. different colour scale to the previous picture).



HadCM3 runs

Standard parameterisation and some of our ensemble members (n.b. different colour scale to the previous picture).





A natural method which is not quite going to work

Imagine that HadCM3 was very fast to run. We could use the following approach:

1. Sample millions of candidates for the collection of simulator parameters, and for each one:
   a. Run the simulator under palaeo-forcing to equilibrium, and
   b. Score the result by comparison with the pseudo-obs.
2. Create a weighted average of the sample.

A natural method which is not quite going to work

Imagine that HadCM3 was very fast to run. We could use the following approach:

1. Sample millions of candidates for the collection of simulator parameters, and for each one:
   a. Run the simulator under palaeo-forcing to equilibrium, and
   b. Score the result by comparison with the pseudo-obs.
2. Create a weighted average of the sample.

Unfortunately for us:

- Each run of HadCM3 takes weeks/months
- We have inherited an ensemble of runs that is not any kind of sample.

The solution is to use the ensemble to construct an emulator of the climate simulator, i.e. a *statistical model of the simulator* that allows us to predict its output at arbitrary parameterisations.

Three steps to an emulator for HadCM3

1 or 2. Consider the simulator $f(r)$ to be the sum of a smooth component $m(r)$ plus internal variability, and estimate $S \approx \mathrm{Var}(\text{internal variability})$.

2 or 1. Dimensionally-reduce the simulator output, keeping only those linear combinations that we trust, D (only a few columns).

3. Estimate the mean and variance functions for the low-dim smooth component, $[D^T m](r)$, using the ensemble and S.

Three steps to an emulator for HadCM3

1 or 2. Consider the simulator $f(r)$ to be the sum of a smooth component $m(r)$ plus internal variability, and estimate $S \approx \mathrm{Var}(\text{internal variability})$.

2 or 1. Dimensionally-reduce the simulator output, keeping only those linear combinations that we trust, D (only a few columns).

3. Estimate the mean and variance functions for the low-dim smooth component, $[D^T m](r)$, using the ensemble and S.

Step 3 produces

$$\text{a mean function } \mu(r) := \mathrm{E}\{[D^T m](r)\}$$
$$\text{and variance function } \Sigma(r) := \mathrm{Var}\{[D^T m](r)\}.$$

Then our emulator for $f(r)$ has mean function $(D^+)^T \mu(r)$ and variance function $(D^+)^T \Sigma(r) D^+$; D^+ is the *Moore-Penrose inverse* of D.

1. Separate out the internal variability

- We think of the simulator as

$$f(r) = m(r) + e(r)$$

where $m(r)$ is a smooth function, and $e(r)$ is a very rough function, representing internal variability.

1. Separate out the internal variability

- We think of the simulator as

$$f(r) = m(r) + e(r)$$

  where $m(r)$ is a smooth function, and $e(r)$ is a very rough function, representing internal variability.

- Now we make a strong assertion in order to proceed:
  - For each r, the simulator has an 'ergodic' attractor, which may vary in location according to r but does not vary (very much) in its gross shape.

  The variance of internal variability is a summary of the shape of the attractor.

1. Separate out the internal variability

- We think of the simulator as

$$f(r) = m(r) + e(r)$$

  where $m(r)$ is a smooth function, and $e(r)$ is a very rough function, representing internal variability.

- Now we make a strong assertion in order to proceed:
  - For each r, the simulator has an 'ergodic' attractor, which may vary in location according to r but does not vary (very much) in its gross shape.

  The variance of internal variability is a summary of the shape of the attractor.

- This strong assertion allows us to estimate the variance of the internal variability at any r, denoted S, using one long 'control run' at the standard setting of the parameters.

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

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

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

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

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

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

- We can, equivalently, write

$$\text{actual climate} \approx (D^+)^T [D^T m](r)$$

where $[D^T m](r)$ is a low-dimensional smooth function.

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

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

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

- We can, equivalently, write

$$\text{actual climate} \approx (D^+)^T [D^T m](r)$$

where $[D^T m](r)$ is a low-dimensional smooth function.

- We are going to emulate $[D^T m](r)$ for arbitrary r. Then we recover actual climate by pre-multiplying by $(D^+)^T$.

(Note that we are 'throwing away' the simulator's internal variability: we do not consider it relevant for reconstructing mean climate.)

3. Emulate $[D^T m](r)$

- Suppose we were to write $[D^T m](r) = B^T r$, where B is a matrix of unknown regression coefficients. Then we would have, starting from $F = M + E$,

$$FD = MD + ED = RB + ED,$$

where R is the matrix of different parameter settings, F the matrix of simulator outputs (one row per run), and E is the matrix of internal variability.

3. Emulate $[D^T m](r)$

- Suppose we were to write $[D^T m](r) = B^T r$, where B is a matrix of unknown regression coefficients. Then we would have, starting from $F = M + E$,

$$FD = MD + ED = RB + ED,$$

where R is the matrix of different parameter settings, F the matrix of simulator outputs (one row per run), and E is the matrix of internal variability.

- Since we know the variance of ED, namely

$$\mathrm{Var}(vec\, ED) = (D^T SD) \otimes I$$

it is a standard calculation to update the mean and variance of $\beta := vec\, B$ using the values R, F, D, and S.

3. Emulate $[D^T m](r)$ (cont)

- Once we have the updated mean and variance of β, then the mean and variance functions for $[D^T m](r)$ follow immediately:

$$\mu(r) = (I \otimes r)^T \, \mathrm{E}_F(\beta)$$
$$\Sigma(r) = (I \otimes r)^T \, \mathrm{Var}_F(\beta)(I \otimes r).$$

  Note that in general $\Sigma(r) > \mathbf{0}$; $\mu(r)$ is *not* claiming to be a perfect surrogate for $m(r)$.

3. Emulate $[D^T m](r)$ (cont)

- Once we have the updated mean and variance of β, then the mean and variance functions for $[D^T m](r)$ follow immediately:

$$\mu(r) = (I \otimes r)^T \, \mathrm{E}_F(\beta)$$
$$\Sigma(r) = (I \otimes r)^T \, \mathrm{Var}_F(\beta)(I \otimes r).$$

  Note that in general $\Sigma(r) > \mathbf{0}$; $\mu(r)$ is *not* claiming to be a perfect surrogate for $m(r)$.

- In practice, we don't use $[D^T m](r) = B^T r$, but

$$[D^T m](r) = B^T g(r),$$

  where $g(r)$ is a more general vector-valued function of r.

- In most cases we can use a 'vague' initial mean and variance for β, namely $\mathrm{Var}(\beta)^{-1} \to \mathbf{0}$.

4. There is no step four

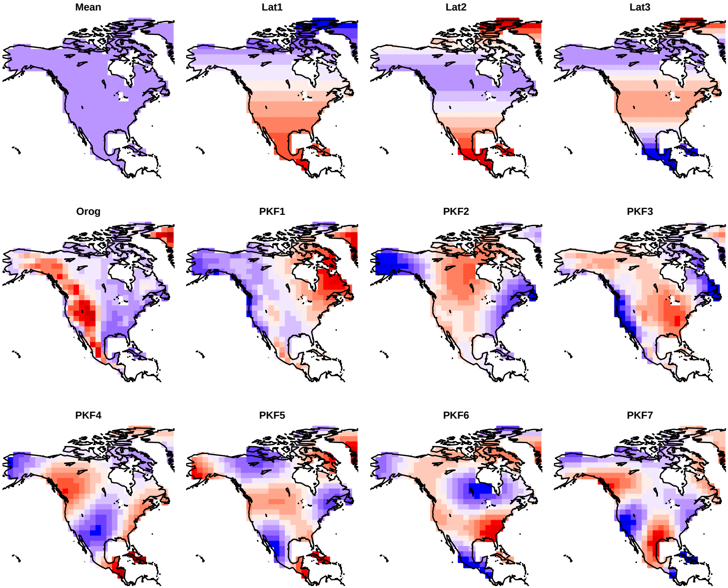
That's it!

We focus our attention on:

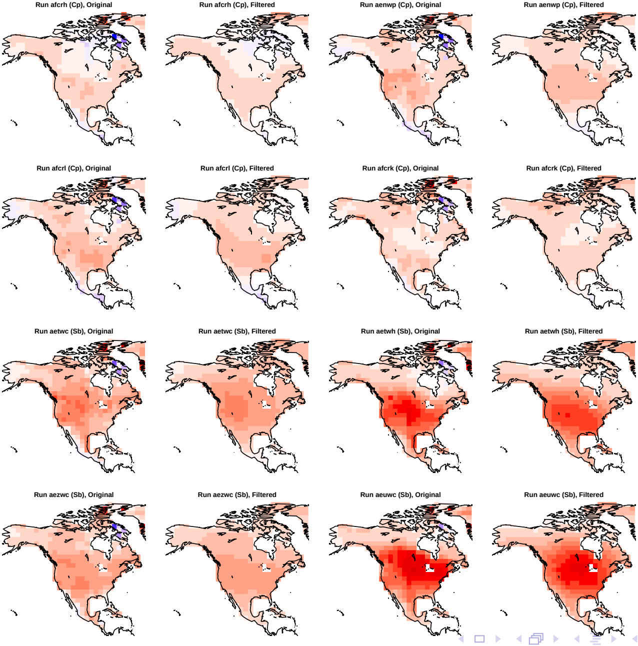
1. Deriving a robust estimate for internal variability;
2. Specifying the reliable linear combinations of the simulator;
3. Choosing regressors to represent the smooth component.

Everything else is just technique.

Our choice of filtering matrix, D

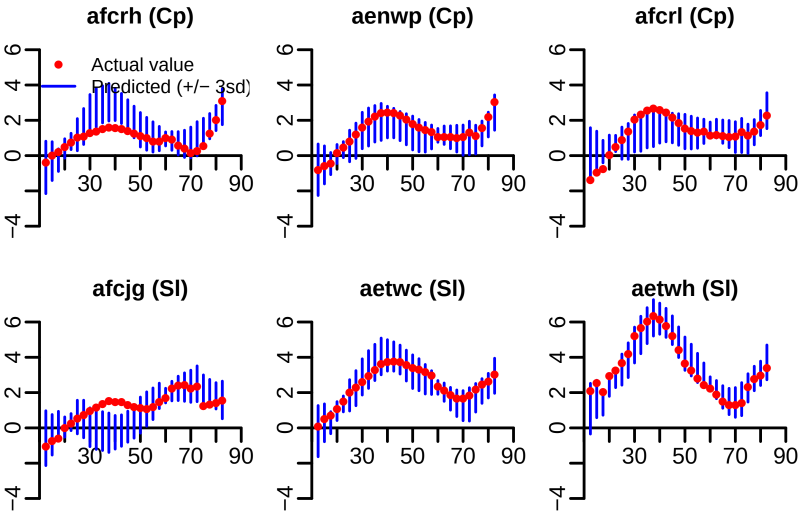


Our choice of filtering matrix, D



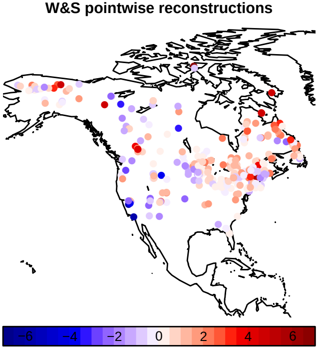
Checking the emulator

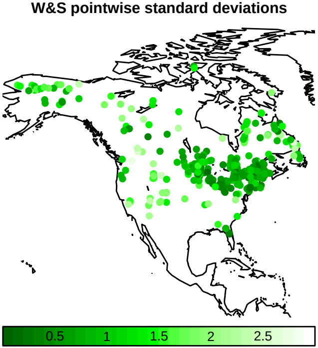
Diagnostic information based on leave-one-out; displayed as zonal means to indicate the emulator's prediction envelope.



Combined reconstruction

Reminder:





Combined reconstruction (cont)

- We separate the parameters into *Control parameters* (e.g. switching between the slab and dynamical ocean) and *Uncertain parameters* (e.g. the entrainment rate in the convection scheme), $r = (r_c, r_u)$.

Combined reconstruction (cont)

- We separate the parameters into *Control parameters* (e.g. switching between the slab and dynamical ocean) and *Uncertain parameters* (e.g. the entrainment rate in the convection scheme), $r = (r_c, r_u)$.
- We link the simulator and the pseudo-obs together in a statistical model:

$$\text{pseudo-obs} = H\underbrace{\left(m(r_c, R_u) + \text{discrepancy}\right)}_{\text{actual climate}} + \text{obs. error}$$

where H is the incidence matrix and $R_u \sim \pi(r_u \mid r_c)$.

Combined reconstruction (cont)

- We separate the parameters into *Control parameters* (e.g. switching between the slab and dynamical ocean) and *Uncertain parameters* (e.g. the entrainment rate in the convection scheme), $r = (r_c, r_u)$.
- We link the simulator and the pseudo-obs together in a statistical model:

$$\text{pseudo-obs} = H \underbrace{\left(m(r_c, R_u) + \text{discrepancy}\right)}_{\text{actual climate}} + \text{obs. error}$$

  where H is the incidence matrix and $R_u \sim \pi(r_u \mid r_c)$.
- We can find the mean and variance of $m(r_c, R_u)$ by *integrating R_u out of the emulator for $m(r)$:*

$$\begin{aligned} \mathrm{E}(m(r_c, R_u)) &= \mathrm{E}(\mu(r_c, R_u)), \text{ and} \\ \mathrm{Var}(m(r_c, R_u)) &= \mathrm{E}(\Sigma(r_c, R_u)) + \mathrm{Var}(\mu(r_c, R_u)) \end{aligned}$$

  (only the first few moments of R_u are likely to be relevant).

Combined reconstruction (cont)

- We separate the parameters into *Control parameters* (e.g. switching between the slab and dynamical ocean) and *Uncertain parameters* (e.g. the entrainment rate in the convection scheme), $r = (r_c, r_u)$.
- We link the simulator and the pseudo-obs together in a statistical model:

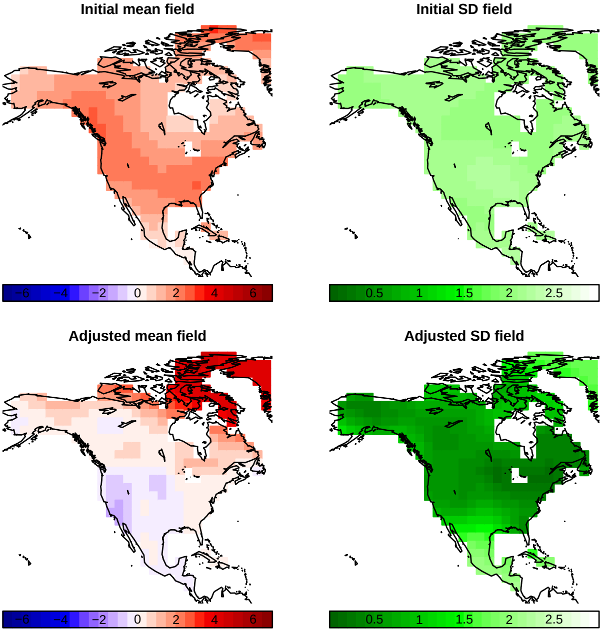
$$\text{pseudo-obs} = H \underbrace{\left(m(r_c, R_u) + \text{discrepancy}\right)}_{\text{actual climate}} + \text{obs. error}$$

  where H is the incidence matrix and $R_u \sim \pi(r_u \mid r_c)$.
- We can find the mean and variance of $m(r_c, R_u)$ by *integrating R_u out of the emulator for $m(r)$:*

$$\begin{aligned} \mathrm{E}(m(r_c, R_u)) &= \mathrm{E}(\mu(r_c, R_u)), \text{ and} \\ \mathrm{Var}(m(r_c, R_u)) &= \mathrm{E}(\Sigma(r_c, R_u)) + \mathrm{Var}(\mu(r_c, R_u)) \end{aligned}$$

  (only the first few moments of R_u are likely to be relevant).
- Now we update the mean and variance of actual climate using the values of the pseudo-obs. We need to specify H, $\pi(r_u \mid r_c)$, $\mathrm{Var}(\text{discrepancy})$, and $\mathrm{Var}(\text{obs. error})$.

Combined reconstruction (cont)



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.

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.

1. Emulating a climate simulator like HadCM3
   - How to get a robust estimate of internal variability?
   - What linear combinations of high-dimensional spatial outputs are 'trustworthy'?
   - How to choose the regression functions for the simulator smooth component?

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.

1. Emulating a climate simulator like HadCM3
   - How to get a robust estimate of internal variability?
   - What linear combinations of high-dimensional spatial outputs are 'trustworthy'?
   - How to choose the regression functions for the simulator smooth component?

2. Linking HadCM3 to reality
   - What is a good probabilistic description for parametric uncertainty?
   - How to assess and quantify structural uncertainty?
   - How to present fully-probabilistic information about spatial (and spatial/temporal) reconstructions?

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.