Calibrating the UVic climate model using principal component emulation

Richard Wilkinson

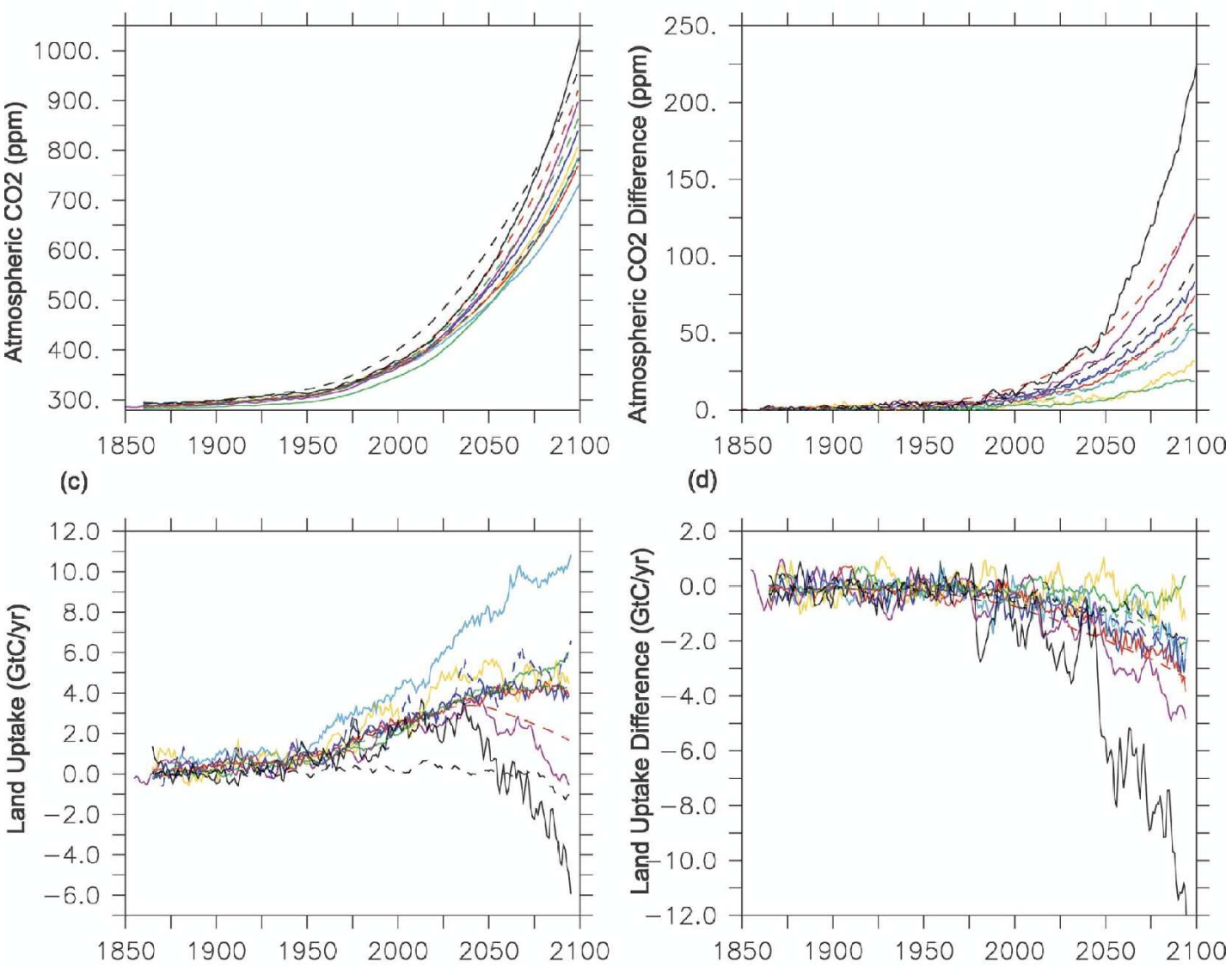
r.d.wilkinson@sheffield.ac.uk
Department of Probability and Statistics
University of Sheffield
MUCM

Joint work with Nathan Urban (Penn. State University)

28 July 2009

Carbon Cycle

Friedlingstein *et al.* 2006 - uncalibrated GCM predictions



Carbon feedbacks

- Terrestrial ecosystems currently absorb a considerable fraction of anthropogenic carbon emissions.
- However, the fate of this sink is highly uncertain due to insufficient knowledge about key feedbacks.
- In particular we are uncertain about the sensitivity of soil respiration to increasing global temperature.
- GCM predictions don't even agree on the sign of the net terrestrial carbon flux.

The figure showed inter-model spread in uncalibrated GCM model predictions.

- How much additional spread is there from parametric (as opposed to model structural) uncertainty?
- Can calibration reduce some of the spread compared to a pile of uncalibrated models?

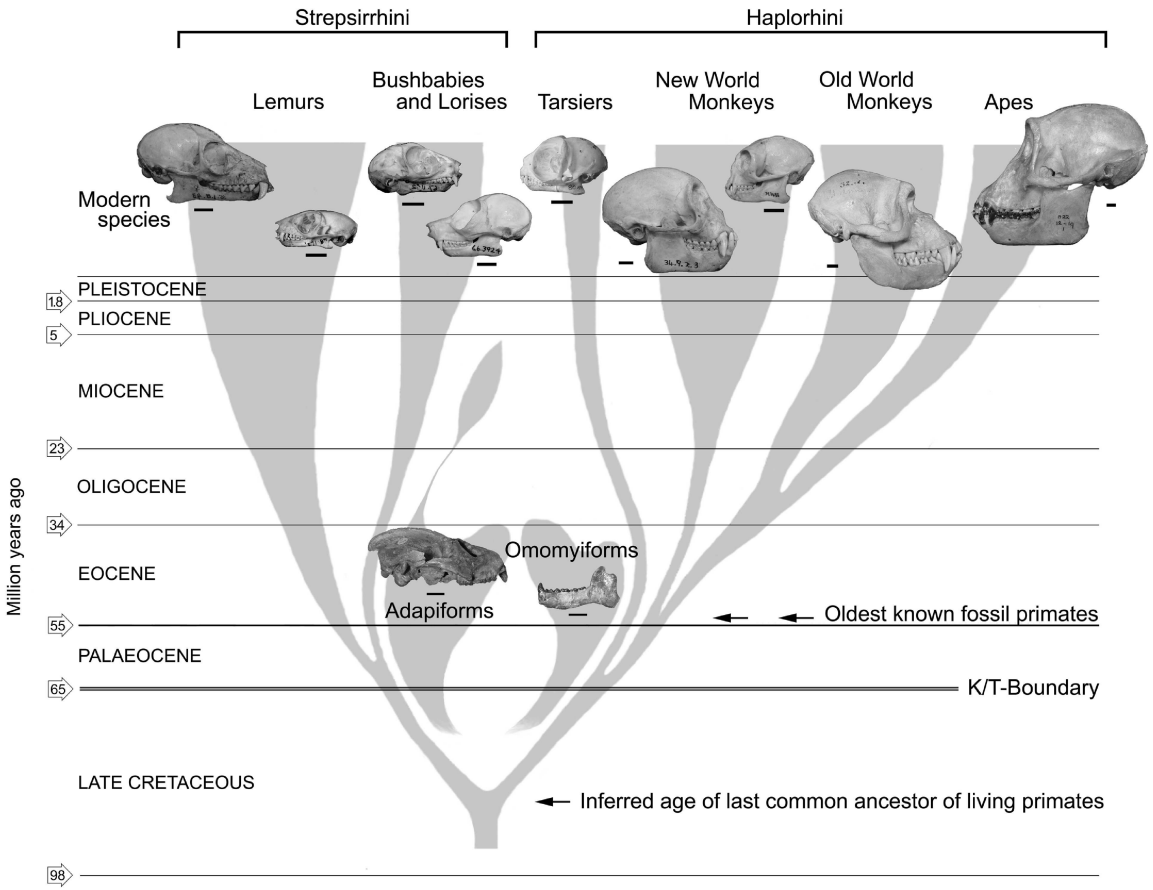
Calibration

The inverse problem

Most models are forwards models, i.e., specify parameters θ and i.c.s and the model $\eta()$ generates output $\mathcal{D}$. Often, we are interested in the inverse-problem, i.e., observe data, want to estimate parameter values.

Different terminology:

- Calibration
- Data assimilation
- Parameter estimation
- Inverse-problem
- Bayesian inference



Computer experiments

Distinguish between two types of input:

- t = control parameters, e.g., time, location, force etc.
- θ = calibration parameters, e.g., gravity, viscosity, respiration sensitivity etc.
  - Physical experiments: nature specifies θ
  - Computer experiments: we must specify θ

Computer experiments

Distinguish between two types of input:

- t = control parameters, e.g., time, location, force etc.
- θ = calibration parameters, e.g., gravity, viscosity, respiration sensitivity etc.
  - Physical experiments: nature specifies θ
  - Computer experiments: we must specify θ

We take the 'best-input' approach:

- θ has a best-fitting value, $\hat{\theta}$, in the sense of representing the data faithfully according to the error structure specified.
- We are not usually ignorant about θ, although $\hat{\theta}$ will not necessarily correspond to true physical values.

Computer experiments

Distinguish between two types of input:

- t = control parameters, e.g., time, location, force etc.
- θ = calibration parameters, e.g., gravity, viscosity, respiration sensitivity etc.
  - Physical experiments: nature specifies θ
  - Computer experiments: we must specify θ

We take the 'best-input' approach:

- θ has a best-fitting value, $\hat{\theta}$, in the sense of representing the data faithfully according to the error structure specified.
- We are not usually ignorant about θ, although $\hat{\theta}$ will not necessarily correspond to true physical values.

Aim: find the posterior distribution of the calibration parameter (θ) given the computer model (η) and the field data ($\mathcal{D}_{\text{field}}$)

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

$$\pi(\theta|\mathcal{D}_{\text{field}}, \eta) \propto \pi(\theta)\mathbb{P}(\mathcal{D}_{\text{field}}|\eta, \theta)$$

UVic Earth System Climate Model

With Nathan Urban (Penn State)

UVic ESCM is an intermediate complexity model with a general circulation ocean and dynamic/thermodynamic sea-ice components coupled to a simple energy/moisture balance atmosphere. It has a dynamic vegetation and terrestrial carbon cycle model (TRIFFID) as well as an inorganic carbon cycle.

- Inputs: Q_{10} = soil respiration sensitivity to temperature (carbon source) and $K_c = CO_2$ fertilization of photosynthesis (carbon sink).
- Output: time-series of CO_2 values, cumulative carbon flux measurements, spatial-temporal field of soil carbon measurements.

UVic Earth System Climate Model

With Nathan Urban (Penn State)

UVic ESCM is an intermediate complexity model with a general circulation ocean and dynamic/thermodynamic sea-ice components coupled to a simple energy/moisture balance atmosphere. It has a dynamic vegetation and terrestrial carbon cycle model (TRIFFID) as well as an inorganic carbon cycle.

- Inputs: Q_{10} = soil respiration sensitivity to temperature (carbon source) and $K_c = CO_2$ fertilization of photosynthesis (carbon sink).
- Output: time-series of CO_2 values, cumulative carbon flux measurements, spatial-temporal field of soil carbon measurements.

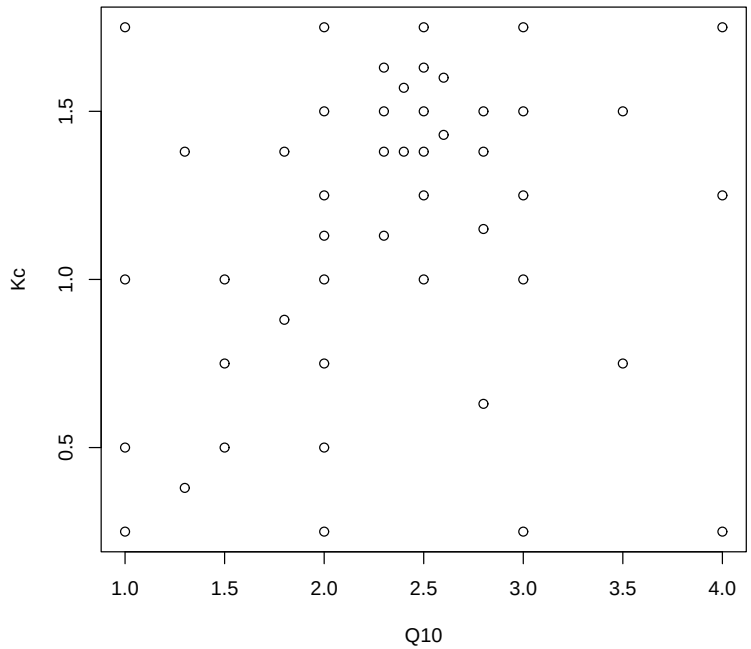
The observational data are limited, and consist of 60 measurements $\mathcal{D}_{field}$:

- 40 instrumental CO_2 measurements from 1960-1999 (from Mauna Loa)
- 17 ice core CO_2 measurements
- 3 cumulative ocean carbon flux measurements

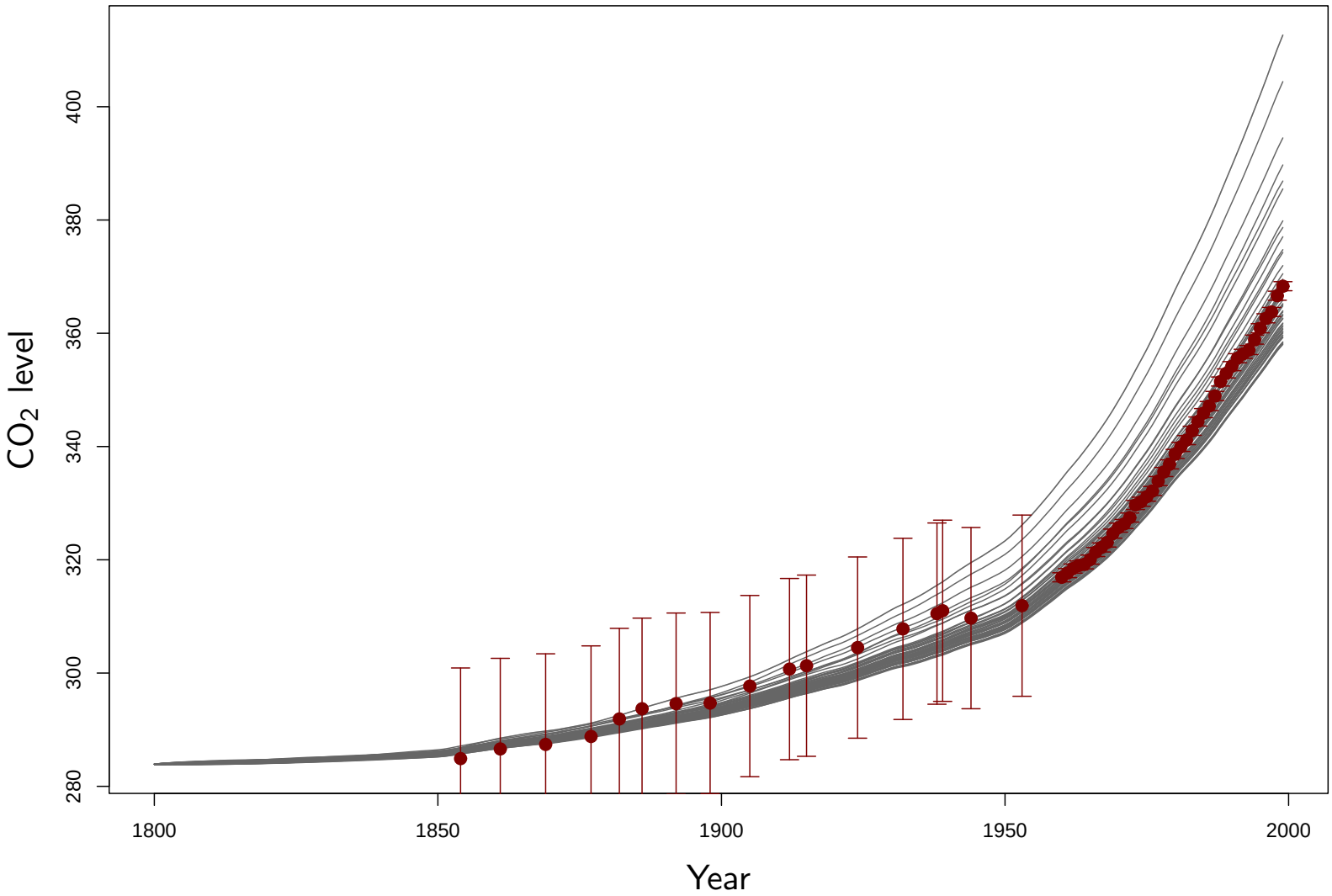
Calibration

The aim is to combine the physics coded into UVic with the empirical observations to learn about the carbon feedbacks.
However, UVic takes approximately two weeks to run for a single input configuration. Consequently, all inference must be done from a limited ensemble of model runs.

- 48 member ensemble, grid design D, output $\mathcal{D}_{sim}$ $(48 \times n)$.



Model runs and data

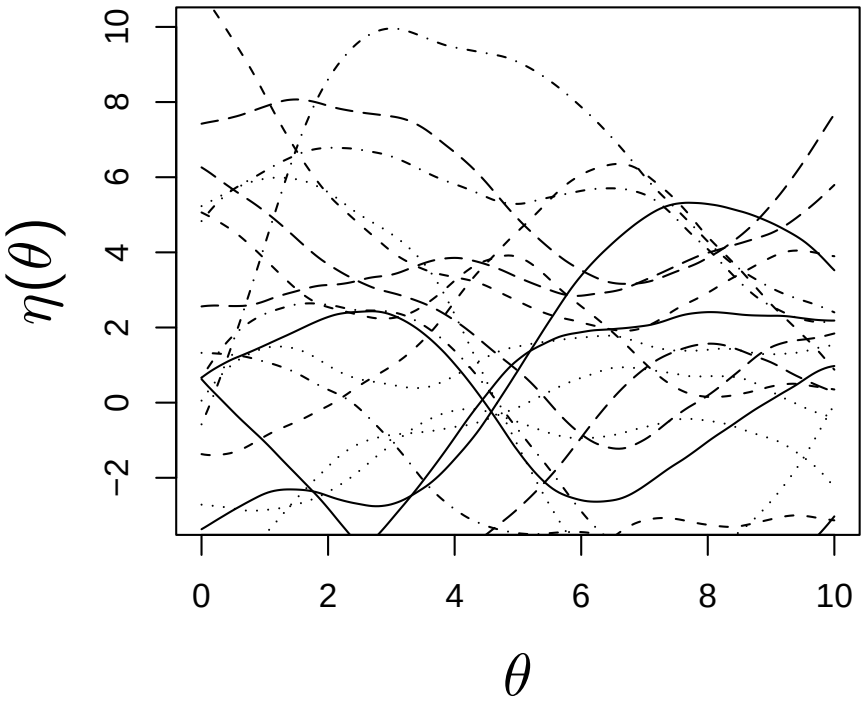


Gaussian Process Emulators

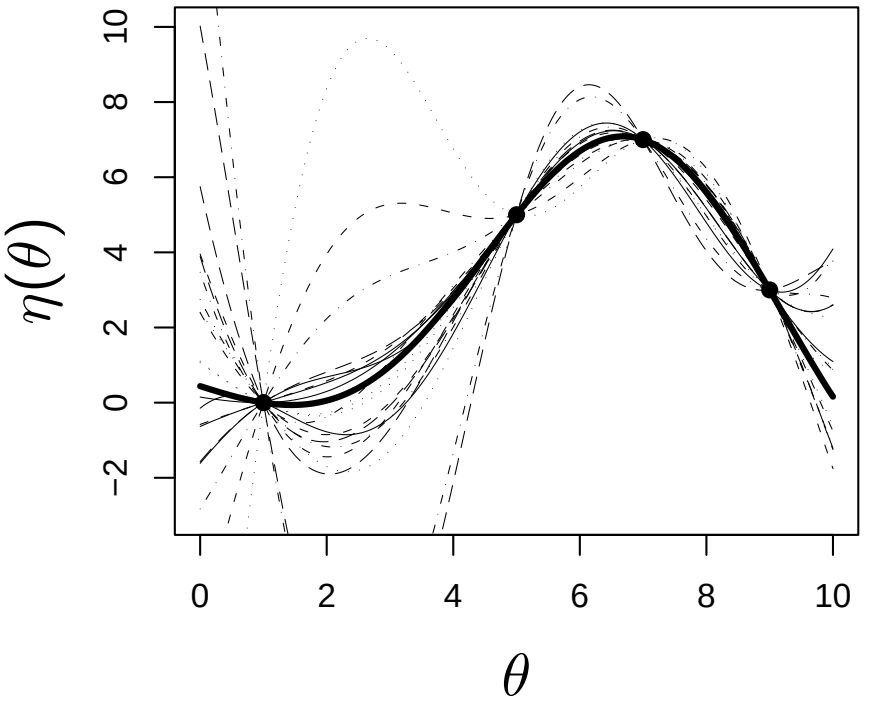
We build emulators (meta-models) to account for code uncertainty

- At untried inputs, we don't know the model's output.
- Assume a priori that $\eta(\cdot) \sim GP(\mu(\cdot), c(\cdot,\cdot))$ for some mean function $\mu(\cdot)$ and covariance function $c(\cdot,\cdot)$, and then condition this on the observed ensemble $\mathcal{D}_{\text{sim}}$.

Unconditioned Gaussian process



Conditioned Gaussian process



Multivariate Emulation

Higdon *et al.* 2008

How can we deal with multivariate ouput?

- Build independent or separable multivariate emulators,
- Outer product emulators,
- Linear model of coregionalization?

Multivariate Emulation

Higdon *et al.* 2008

How can we deal with multivariate ouput?

- Build independent or separable multivariate emulators,
- Outer product emulators,
- Linear model of coregionalization?

Instead, if the outputs are highly correlated we can reduce the dimension of the data by projecting the data onto some lower dimensional manifold $\mathcal{Y}^{pc}$.

Multivariate Emulation

Higdon *et al.* 2008

How can we deal with multivariate ouput?

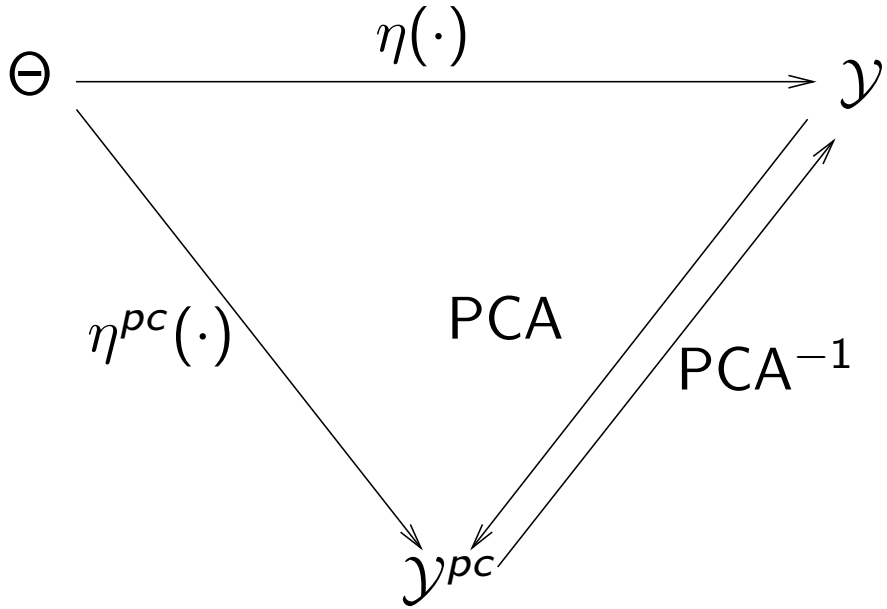
- Build independent or separable multivariate emulators,
- Outer product emulators,
- Linear model of coregionalization?

Instead, if the outputs are highly correlated we can reduce the dimension of the data by projecting the data onto some lower dimensional manifold $\mathcal{Y}^{pc}$.

We can use any dimension reduction technique as long as

- we can reconstruct to the original output space
- we can quantify the reconstruction error.

We can then emulate the function that maps the input space Θ to the reduced dimensional output space $\mathcal{Y}^{pc}$, i.e., $\eta_{pc}(\cdot) : \Theta \rightarrow \mathcal{Y}^{pc}$



Principal Component Emulation (EOF)

We use principal component analysis to project the data onto a lower dimensional manifold, as it is the optimal linear projection (in terms of minimizing reconstruction error).

1. Centre and scale $\mathcal{D}_{\text{sim}}$ so that each column has mean 0 and variance 1. Scaling the columns makes specification of prior distributions for the emulators simpler.
2. Find the singular value decomposition of $\mathcal{D}_{\text{sim}}$.

$$\mathcal{D}_{\text{sim}} = U\Gamma V^*.$$

   Γ contains the singular values (eigenvalues), and V the principal components (eigenvectors).
3. Decide on the dimension of the principal subspace, n^* say, and throw away all but the n^* leading principal components. An orthonormal basis for the principal subspace is given by the first n^* columns of V, denoted V_1. Let V_2 be the matrix of discarded columns.
4. Project $\mathcal{D}_{\text{sim}}$ onto the principal subspace to find $\mathcal{D}^{pc}_{\text{sim}} = \mathcal{D}_{\text{sim}} V_1$

PCA emulation

We then emulate the reduced dimension model

$$\eta_{pc}(\cdot) = (\eta_{pc}^1(\cdot), \ldots, \eta_{pc}^{n^*}(\cdot)).$$

- Each component η_{pc}^i will be uncorrelated (in the ensemble) but not necessarily independent. We use independent Gaussian processes for each component, which seems to be an adequate approximation in all the examples we've looked at.
- The output can be reconstructed from the principal component space to the original full space, accounting for reconstruction error, by a simple matrix multiplication and modelling the discarded components as Gaussian noise with variance equal to the corresponding eigenvalue:

$$\eta(\theta) = V_1\eta_{pc}(\theta) + V_2\text{diag}(\Lambda)$$

where $\Lambda_i \sim N(0, \Gamma_{ii})$ ($\Gamma_{ii} = i^{th}$ eigenvalue).

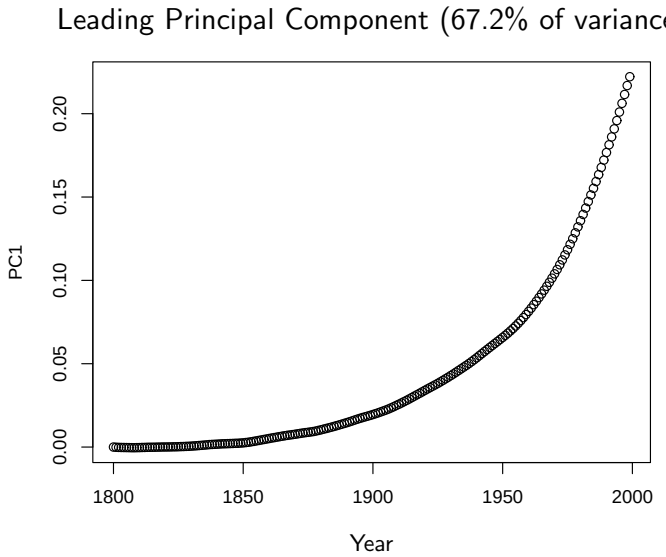
Comments

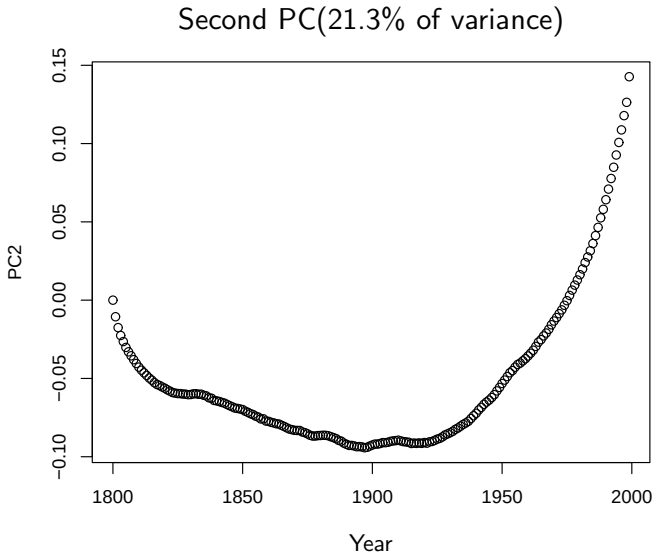
- This approach (PCA emulation) requires that the outputs are highly correlated.
- We are assuming that the output $\mathcal{D}_{\text{sim}}$ is really a linear combination of a smaller number of variables,

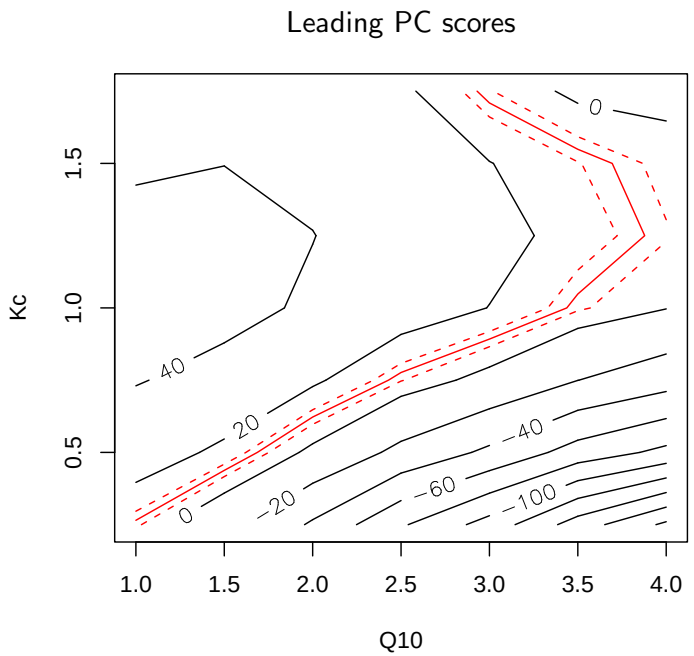
$$\eta(\theta) = \mathbf{v}_1 \eta_{pc}^1(\theta) + \ldots + \mathbf{v}_{n^*} \eta_{pc}^{n^*}(\theta)$$

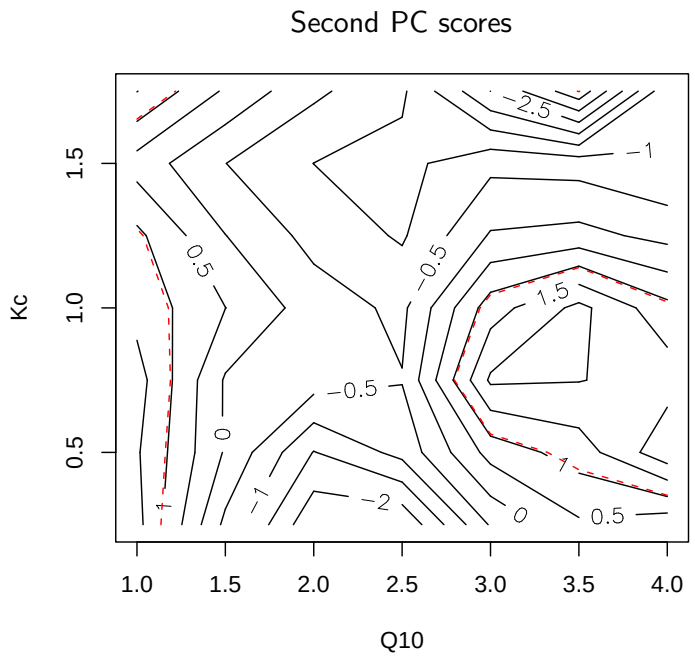
  which may be a reasonable assumption in many situations, eg, temporal spatial fields.
- Although PCA is a linear method, the method can be used on highly non-linear models as we are still using non-linear Gaussian processes to map from Θ to $\mathcal{Y}^{pc}$.
- This method accounts for code uncertainty and automatically accounts for the reconstruction error caused by reducing the dimension of the data.

PC Plots


Leading Principal Component (67.2% of variance)
PC1
0.00
0.05
0.10
0.15
0.20
1800
1850
1900
1950
2000
Year


Second PC(21.3% of variance)
PC2
−0.10
−0.05
0.00
0.05
0.10
0.15
1800
1850
1900
1950
2000
Year


Leading PC scores
Kc
0.5
1.0
1.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Q10
0
40
20
−20
−40
−60
−100


Second PC scores
Kc
0.5
1.0
1.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Q10
2.5
1
0.5
−0.5
1.5
−0.5
0
−1
−2
0
0.5
1

GP Choices

Choice of regressors

- We use products of Legendre polynomials on $[-1, 1]$ (Rougier 2007) - an orthonormal basis. We allow up to quadratic terms

GP Choices

Choice of regressors

- We use products of Legendre polynomials on $[-1, 1]$ (Rougier 2007) - an orthonormal basis. We allow up to quadratic terms

Covariance function

- Matern with $\nu = 5/2 \Rightarrow$ twice differentiable output

$$c_{5/2}(r) = \tau \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right) \exp\left(-\frac{\sqrt{5}r}{l}\right)$$

- Give τ a $\Gamma(1.5, 6)$ prior distribution in all the principal component emulators.

GP Choices

Choice of regressors

- We use products of Legendre polynomials on $[-1, 1]$ (Rougier 2007) - an orthonormal basis. We allow up to quadratic terms

Covariance function

- Matern with $\nu = 5/2 \Rightarrow$ twice differentiable output

$$c_{5/2}(r) = \tau \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right) \exp\left(-\frac{\sqrt{5}r}{l}\right)$$

- Give τ a $\Gamma(1.5, 6)$ prior distribution in all the principal component emulators.

Length scales l

- Estimate and fix the length scales using their maximum likelihood estimates.

GP Choices

Choice of regressors

- We use products of Legendre polynomials on $[-1, 1]$ (Rougier 2007) - an orthonormal basis. We allow up to quadratic terms

Covariance function

- Matern with $\nu = 5/2 \Rightarrow$ twice differentiable output

$$c_{5/2}(r) = \tau \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right) \exp\left(-\frac{\sqrt{5}r}{l}\right)$$

- Give τ a $\Gamma(1.5, 6)$ prior distribution in all the principal component emulators.

Length scales l

- Estimate and fix the length scales using their maximum likelihood estimates.

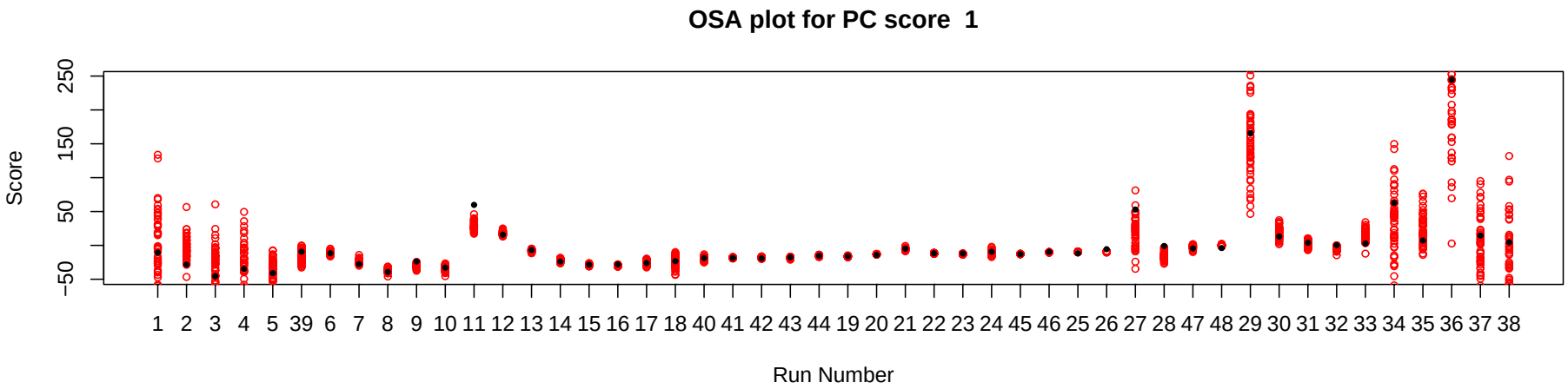
Alternative:

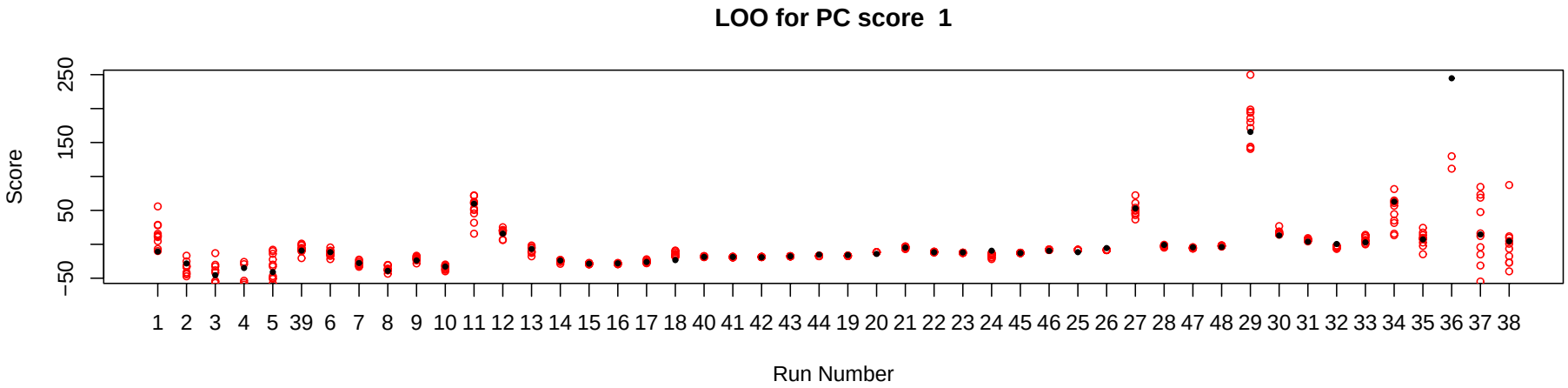
- τ and l are often not both identifiable. Instead, fix l using Addler's theorem by considering the expected number of up-crossings by the residual.

Diagnostics

One-step-ahead (OSA) and leave-one-out (LOA) for PC1

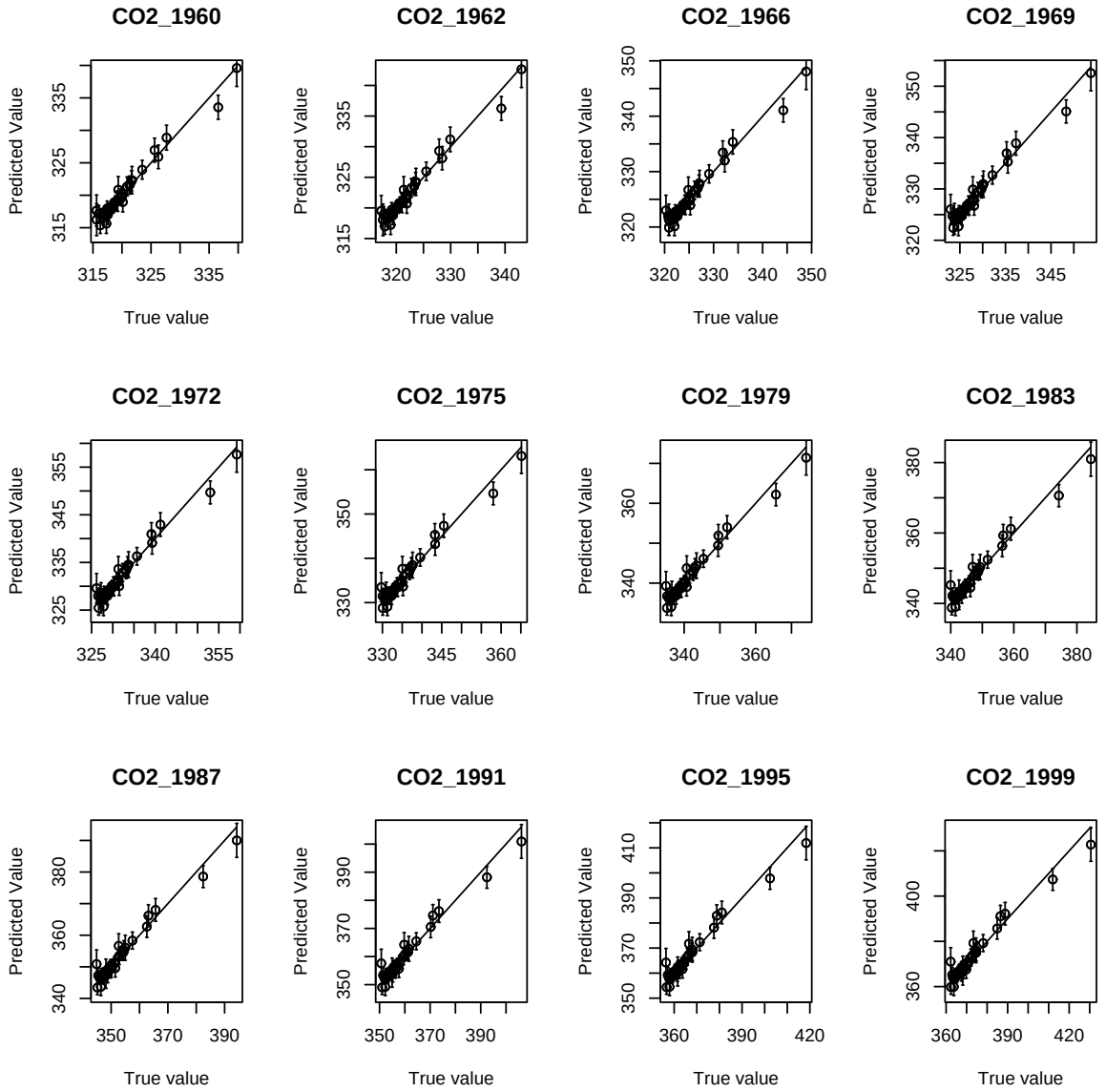
Order the ensemble according to θ_1.





Emulator Diagnostics

LOO cross-validation plots, using 10 PCs (99.2% of variance explained)



Calibration Framework

Kennedy and O'Hagan 2001

We have two sources of information:

- Computer model $\eta(t, \theta)$
  - with a limited ensemble of model runs $\mathcal{D}_{sim} = \{\eta(t_i, \theta_i), i = \ldots\}$.
- Field data $\mathcal{D}_{\text{field}}$: a collection of noisy measurements of reality at a variety of t values.

Calibration Framework

Kennedy and O'Hagan 2001

We have two sources of information:

- Computer model $\eta(t, \theta)$
  - with a limited ensemble of model runs $\mathcal{D}_{sim} = \{\eta(t_i, \theta_i), i = \ldots\}$.
- Field data $\mathcal{D}_{\text{field}}$: a collection of noisy measurements of reality at a variety of t values.

Many assimilation approaches assume that measurements represent the *computer model* run at its best input value plus independent random noise. If the model is wrong, this is assumption is false. At best, we observe *reality* plus independent random noise.

Calibration Framework

Kennedy and O'Hagan 2001

We have two sources of information:

- Computer model $\eta(t, \theta)$
  - with a limited ensemble of model runs $\mathcal{D}_{sim} = \{\eta(t_i, \theta_i), i = \ldots\}$.
- Field data $\mathcal{D}_{\text{field}}$: a collection of noisy measurements of reality at a variety of t values.

Many assimilation approaches assume that measurements represent the *computer model* run at its best input value plus independent random noise. If the model is wrong, this is assumption is false. At best, we observe *reality* plus independent random noise.
Instead, include an additional model error term.

- Measurement error ϵ
- Model discrepancy $\delta(t)$

Calibration Framework

Assume that reality $\zeta(t)$ is the computer model run at the 'true' value of the parameter $\hat{\theta}$ plus model error:

$$\zeta(t) = \eta(t, \hat{\theta}) + \delta(t)$$

Calibration Framework

Assume that reality $\zeta(t)$ is the computer model run at the 'true' value of the parameter $\hat{\theta}$ plus model error:

$$\zeta(t) = \eta(t, \hat{\theta}) + \delta(t)$$

We observe reality plus noise:

$$\mathcal{D}_{field}(t) = \zeta(t) + \epsilon(t)$$

so that

$$\mathcal{D}_{field}(t) = \eta(t, \hat{\theta}) + \delta(t) + \epsilon(t).$$

Calibration Framework

Assume that reality $\zeta(t)$ is the computer model run at the 'true' value of the parameter $\hat{\theta}$ plus model error:

$$\zeta(t) = \eta(t, \hat{\theta}) + \delta(t)$$

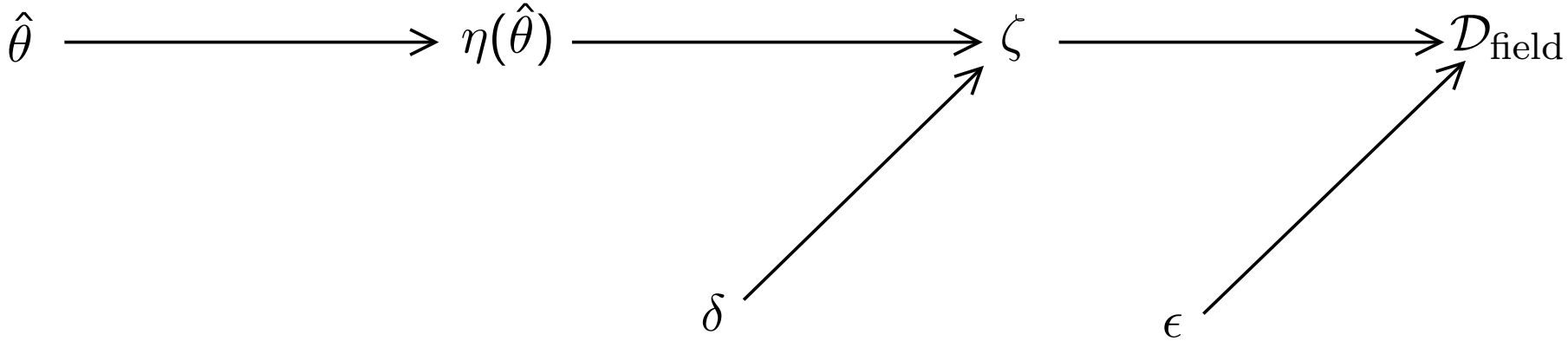
We observe reality plus noise:

$$\mathcal{D}_{field}(t) = \zeta(t) + \epsilon(t)$$

so that

$$\mathcal{D}_{field}(t) = \eta(t, \hat{\theta}) + \delta(t) + \epsilon(t).$$

We then aim to find $\pi(\hat{\theta}|\mathcal{D}_{sim}, \mathcal{D}_{field})$.



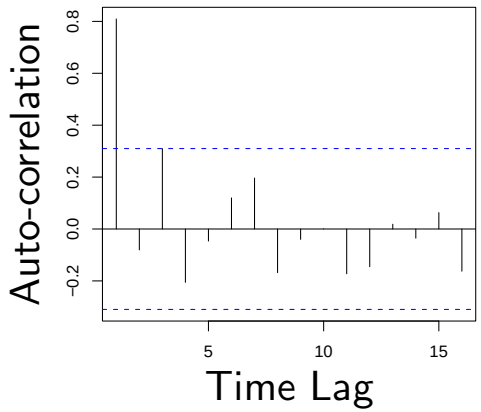
Model Discrepancy

The calibration framework used is:

$$\mathcal{D}_{field}(t) = \eta(\theta, t) + \delta(t) + \epsilon(t)$$

The model predicts the underlying trend, but real climate fluctuates around this. We model

- discrepancy as an AR1 process: $\delta(0) \sim N(0, \sigma_\delta^2)$, and $\delta(t) = \rho\delta(t-1) + N(0, \sigma_\delta^2)$.
- Measurement error as heteroscedastic independent random noise $\epsilon(t) \sim N(0, \lambda(t))$.



How should we better model this discrepancy?

MCMC

Metropolis-within-Gibbs Sampler

Prior distributions: $\rho \sim \Gamma(5, 1)$, $\sigma_\delta^2 \sim \Gamma(4, 0.6)$, $\sigma^2 \sim \Gamma(1.5, 6)$, $\theta = (Q_{10}, K_c)$, $Q_{10} \sim U[1, 4]$, $K_c \sim U[0.25, 1.75]$.

MCMC

Metropolis-within-Gibbs Sampler

Prior distributions: $\rho \sim \Gamma(5,1)$, $\sigma_\delta^2 \sim \Gamma(4,0.6)$, $\sigma^2 \sim \Gamma(1.5,6)$, $\theta = (Q_{10}, K_c)$, $Q_{10} \sim U[1,4]$, $K_c \sim U[0.25,1.75]$.
We can then use a Metropolis-within-Gibbs sampler to find the posterior distribution

$$\pi(\theta, \sigma^2, \rho, \sigma_\delta^2 | \mathcal{D}_{\text{field}}, \mathcal{D}_{\text{sim}})$$

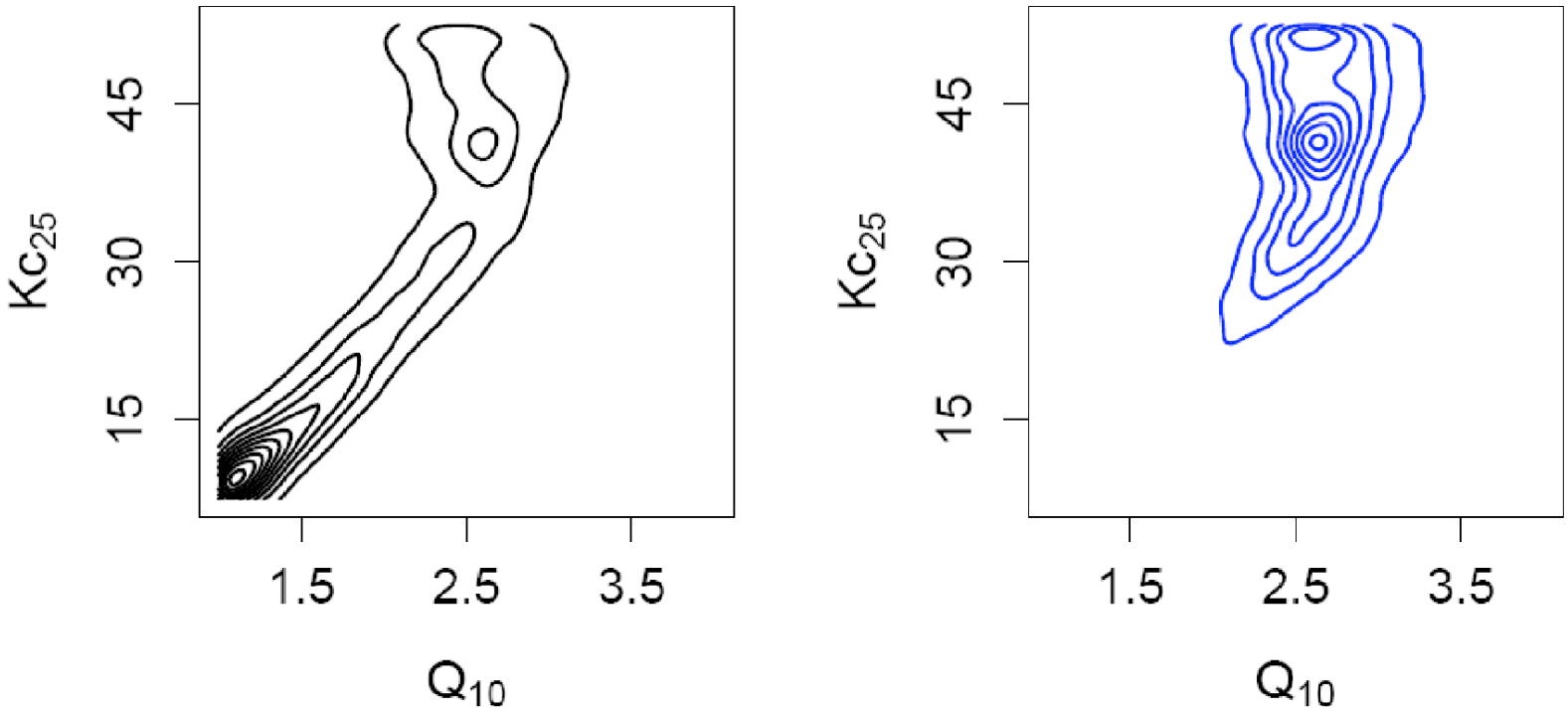
using the following steps

$$\pi(\sigma^2 | \theta, \rho, \sigma_\delta^2, \mathcal{D}_{\text{sim}}, \mathcal{D}_{\text{field}}) \text{ - Gibbs update}$$
$$\pi(\theta | \sigma^2, \rho, \sigma_\delta^2, \mathcal{D}_{\text{sim}}, \mathcal{D}_{\text{field}}) \text{ - MH step}$$
$$\pi(\rho, \sigma_\delta^2 | \theta, \sigma^2, \mathcal{D}_{\text{sim}}, \mathcal{D}_{\text{field}}) \text{ - MH step}$$

Reparameterizing in terms of $\log(\rho)$ and using a block update for ρ and σ_δ^2 helps with the convergence.

Results

Posterior distributions when using uniform prior distributions (left plot) for both parameters, and when using an observation based prior for Q_{10} (right plot).



At low K_c there is positive correlation between K_c and Q_{10}, but this reverses to negative correlation at high K_c - a result of non-linearities in the response of carbon fertilization to CO_2 and respiration to temperature.

Conclusions

- For highly correlated multivariate output principal component emulation can work well and is computationally cheap and easy to implement.
- A large number of output dimensions can be reduced to a smaller number of principal component scores which can then be emulated, accounting for any error in the compression.
- Given the model, forcing data, constraints and uniform priors, high values of Q_{10} are excluded but no value of K_c can be ruled out.
- Acceptable parameter combinations produce similar responses of the carbon cycle during the years 1800-1999 but produce widely divergent future predictions.