# Inference for complex models

Richard Wilkinson

School of Mathematical Sciences
University of Nottingham

23 April 2015

# Computer experiments

Rohrlich (1991): Computer simulation is

> *'a key milestone somewhat comparable to the milestone that started the empirical approach (Galileo) and the deterministic mathematical approach to dynamics (Newton and Laplace)'*

Challenges for statistics:
How do we make inferences about the world from a simulation of it?

# Computer experiments

Rohrlich (1991): Computer simulation is

> *'a key milestone somewhat comparable to the milestone that started the empirical approach (Galileo) and the deterministic mathematical approach to dynamics (Newton and Laplace)'*

Challenges for statistics:
How do we make inferences about the world from a simulation of it?

- how do we relate simulators to reality?
- how do we estimate tunable parameters?
- how do we deal with computational constraints?
- how do we make uncertainty statements about the world that combine models, data and their corresponding errors?

There is an inherent a lack of quantitative information on the uncertainty surrounding a simulation - unlike in physical experiments.

# Bayesian statistics

Represent all uncertainties as probability distributions:

$$\pi(\theta|D) = \frac{\pi(D|\theta)\pi(\theta)}{\pi(D)}$$

# Bayesian statistics

Represent all uncertainties as probability distributions:

$$\pi(\theta|D) = \frac{\pi(D|\theta)\pi(\theta)}{\pi(D)}$$

- $\pi(\theta|D)$ is the posterior distribution
  - Always hard to compute: SMC$^2$, PGAS, Tempered NUTS-HMC

# Bayesian statistics

Represent all uncertainties as probability distributions:

$$\pi(\theta|D) = \frac{\pi(D|\theta)\pi(\theta)}{\pi(D)}$$

- $\pi(\theta|D)$ is the posterior distribution
  - Always hard to compute: SMC$^2$, PGAS, Tempered NUTS-HMC
- $\pi(D|\theta)$ is the likelihood function.
  - For complex models can be slow to compute: GP emulators
  - Can also be impossible to compute in some cases: ABC
  - $$\pi(D|\theta) = \int \pi(D|X)\pi(X|\theta)\mathrm{d}X$$

    Relating simulator to reality can make specifying $\pi(D|\theta)$ particularly difficult: Simlator discrepancy modelling

# Bayesian statistics

Represent all uncertainties as probability distributions:

$$\pi(\theta|D) = \frac{\pi(D|\theta)\pi(\theta)}{\pi(D)}$$

- $\pi(\theta|D)$ is the posterior distribution
  - Always hard to compute: SMC$^2$, PGAS, Tempered NUTS-HMC
- $\pi(D|\theta)$ is the likelihood function.
  - For complex models can be slow to compute: GP emulators
  - Can also be impossible to compute in some cases: ABC
  - $$\pi(D|\theta) = \int \pi(D|X)\pi(X|\theta)\mathrm{d}X$$

    Relating simulator to reality can make specifying $\pi(D|\theta)$ particularly difficult: Simlator discrepancy modelling
- $\pi(D)$ is the model evidence or normalising constant.
  - Requires us to integrate, and is thus harder to compute than $\pi(\theta|D)$: SMC$^2$, nested sampling

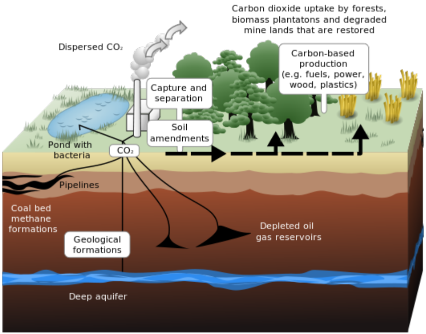
# Outline

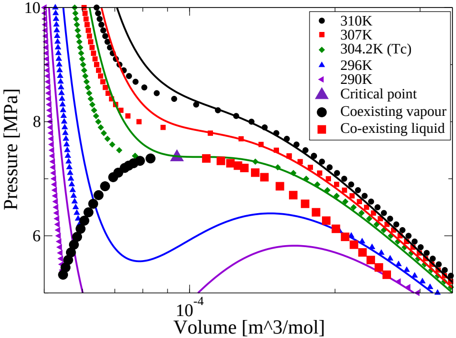
Use scientific need to motivate methodological work.

1. Carbon capture and storage: estimating storage properties
   - Gaussian process (GP) emulators
   - Dealing with high dimension and physical constraints
2. Population genetics: inferring divergence times
   - ABC
   - Accelerating inference using GP-ABC
3. Climate science: palaeoclimate reconstruction
   - Model selection
   - simulator discrepancy

# Uncertainty quantification for Carbon Capture and Storage

## EPSRC: transport




Technical challenges:

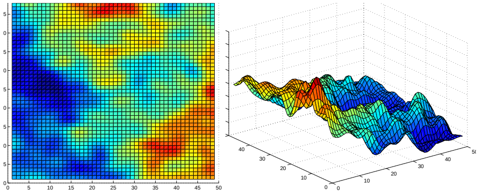
- How do we find non-parametric Gaussian process models that i) obey the fugacity constraints ii) have the correct asymptotic behaviour
- How do we fit parametric equations of state (Peng-Robinson and variants) - tempered NUTS-HMC.

# Storage Panacea

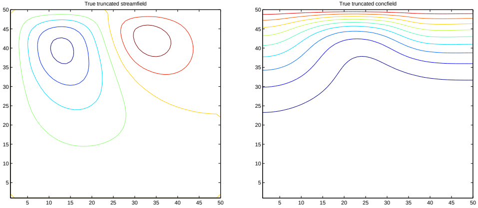
Knowledge of the physical problem is encoded in a simulator $f$

Inputs:

Permeability field, K
(2d field)



$\downarrow f(K)$

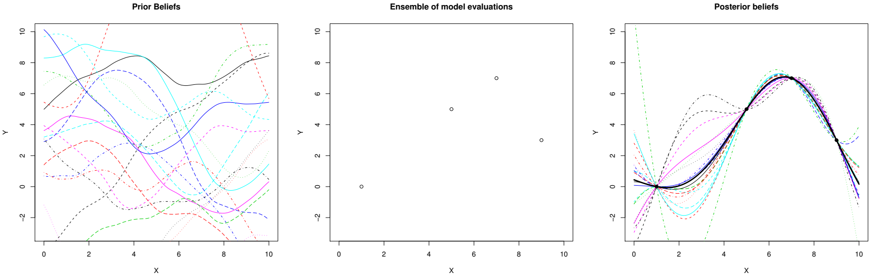



Outputs:

Stream func. (2d field),
concentration (2d field),
surface flux (1d scalar),
$\vdots$

Surface Flux= 6.43, . . .

# Gaussian process emulators




How do we deal with high dimensional inputs and outputs?

- Represent input fields by truncated Karhunen-Loeve expansions
- Represent output fields by truncated principal component expansions.

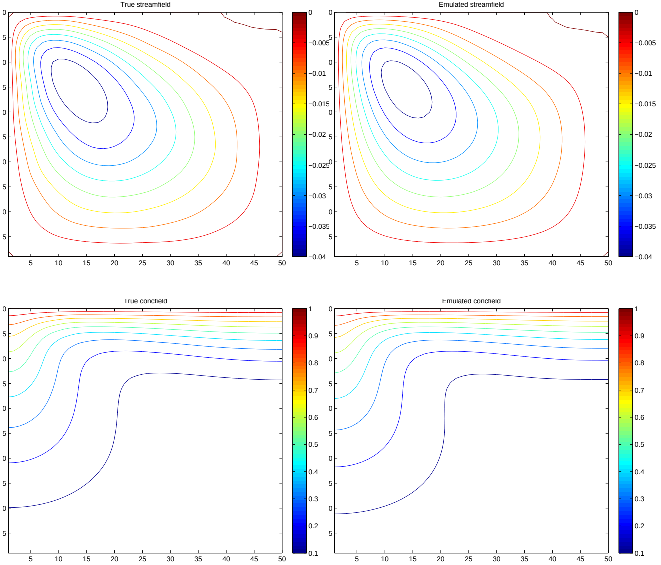
Challenges:

- Design
- Nugget issues

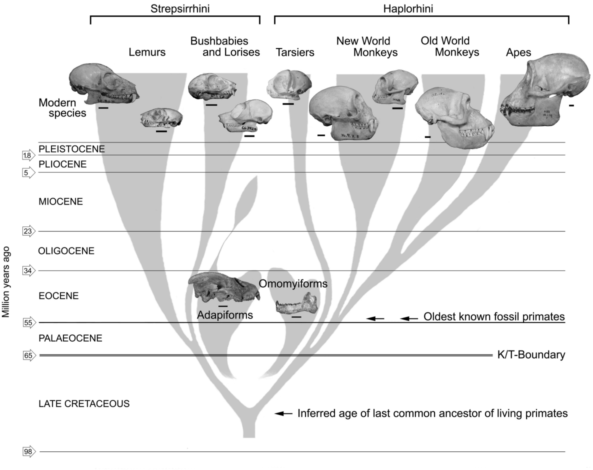
Wilkinson 2010, Holden *et al.* 2015, Bonceur *et al.* 2015.

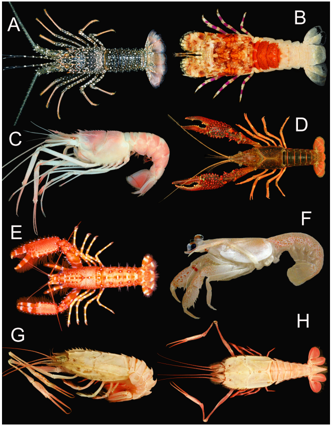
# CCS examples

## Left=true, right = emulated, 118 training runs, held out test set.



# Estimating Divergence Times

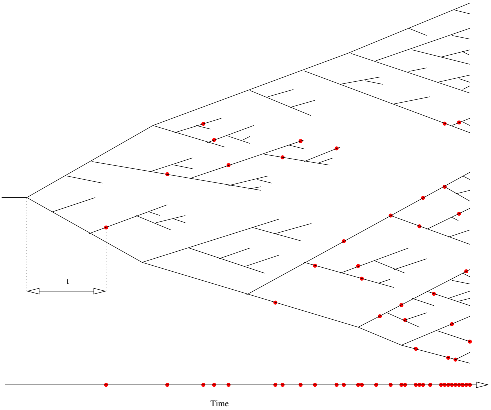


# Forward simulation

Model evolution and fossil finds

- Let $\tau$ be the temporal gap between the divergence time and the oldest fossil.




The posterior for $\tau$ is then used as a prior for a genetic analysis.

The likelihood function $\pi(D|\theta)$ is intractable, but it is cheap to simulate.

# Approximate Bayesian Computation (ABC)

Wilkinson 2008/2013, Wilkinson and Tavaré 2009

If the likelihood function is intractable, then ABC is one of the few approaches we can use to do inference.

### Uniform Rejection Algorithm

- Draw $\theta$ from $\pi(\theta)$
- Simulate $X \sim f(\theta)$
- Accept $\theta$ if $\rho(D, X) \leq \epsilon$

$\epsilon$ reflects the tension between computability and accuracy.

- As $\epsilon \to \infty$, we get observations from the prior, $\pi(\theta)$.
- If $\epsilon = 0$, we generate observations from $\pi(\theta \mid D)$.

# Approximate Bayesian Computation (ABC)

Wilkinson 2008/2013, Wilkinson and Tavaré 2009

If the likelihood function is intractable, then ABC is one of the few approaches we can use to do inference.

## Uniform Rejection Algorithm

- Draw $\theta$ from $\pi(\theta)$
- Simulate $X \sim f(\theta)$
- Accept $\theta$ if $\rho(D, X) \leq \epsilon$

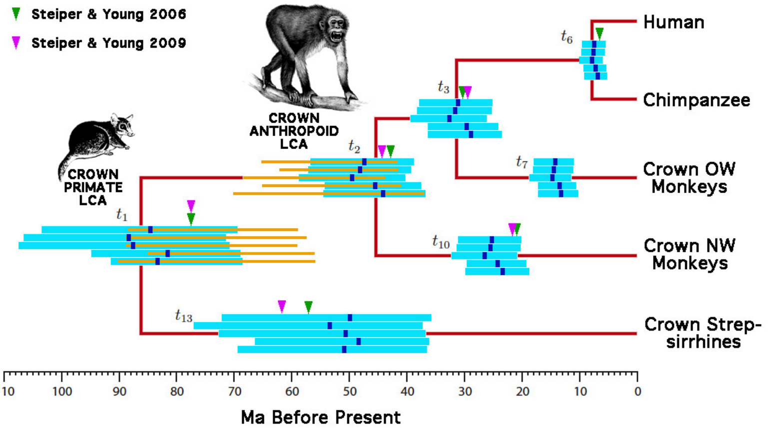
$\epsilon$ reflects the tension between computability and accuracy.

- As $\epsilon \to \infty$, we get observations from the prior, $\pi(\theta)$.
- If $\epsilon = 0$, we generate observations from $\pi(\theta \mid D)$.

ABC does not require explicit knowledge of the likelihood function

- Inference is done using simulation from the model
- Very popular in biological disciplines
  - Simple to implement
  - Intuitive
  - Can usually be applied

# An integrated molecular and palaeontological analysis




The fossil record does not constrain the primate divergence time as closely as previously believed.

- Genetic and palaeontology estimates unified
- Human-chimp divergence time pushed further back.

Wilkinson *et al.* 2011, Bracken-Grissom *et al.* 2014.

# Accelerating ABC: GP-ABC

Monte Carlo methods (such as ABC) are costly and can require more simulation than is possible. However,

- most methods sample naively - they don't learn from previous simulations
- they don't exploit known properties of the likelihood function, such as continuity
- they sample randomly, rather than using careful design.

Emulators are the usual approach to dealing with complex models. But, emulating stochastic simulators is problematic.

# Accelerating ABC: GP-ABC

Monte Carlo methods (such as ABC) are costly and can require more simulation than is possible. However,

- most methods sample naively - they don't learn from previous simulations
- they don't exploit known properties of the likelihood function, such as continuity
- they sample randomly, rather than using careful design.

Emulators are the usual approach to dealing with complex models. But, emulating stochastic simulators is problematic.

Instead of modelling the simulator output, we can instead model $L(\theta) = \pi(D|\theta)$

- $D$ remains fixed: we only need learn $L$ as a function of $\theta$
- 1d response surface
- But, it can be hard to model.

# History matching waves

The log-likelihood for a typical problem ranges over several orders of magnitude.

Consequently, it is hard to model the log-likelihood across the entire parameter space.

# History matching waves

The log-likelihood for a typical problem ranges over several orders of magnitude.

Consequently, it is hard to model the log-likelihood across the entire parameter space.

- Introduce waves of history matching.
- In each wave, build a GP model that can rule out regions of space as implausible.

Using GP-ABC it is possible to obtain the same results as more complex MC algorithms, but with two-orders of magnitude fewer simulations (Wilkinson 2014, Holden *et al.* 2015 (in revision)).

# Active learning for history-matching/GP-ABC

## Work with James Hensman

Sequential design is the key to further reducing computational burden.

- Given our current knowledge, where should we next run the simulator to most improve our knowledge?

One option is to minimise the expected average entropy of the history match

# Active learning for history-matching/GP-ABC

## Work with James Hensman

Sequential design is the key to further reducing computational burden.

- Given our current knowledge, where should we next run the simulator to most improve our knowledge?

One option is to minimise the expected average entropy of the history match

- Any GP emulator allows us to calculate a probabilistic classification

$$p(\theta) = \mathbb{P}(\theta \text{ implausible})$$

- The entropy of our belief at $\theta$ is

$$E(\theta) = -p \log p - (1-p)\log(1-p)$$

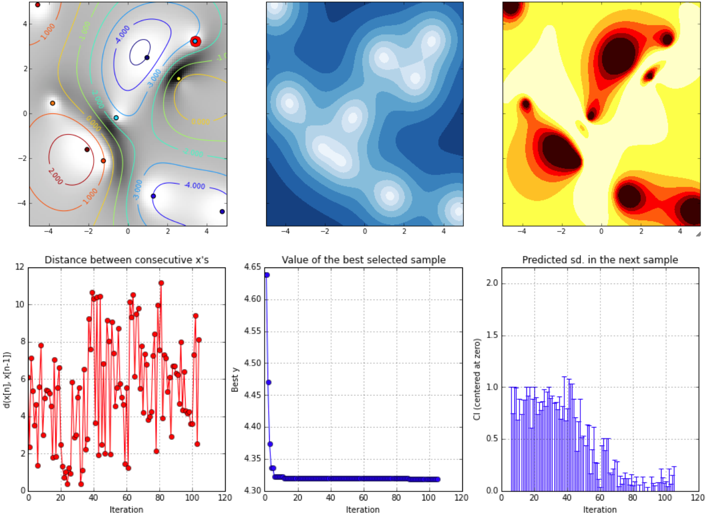
  and the average entropy is

$$\mathcal{E} = \int E(\theta)\mathrm{d}\theta$$

- Choose the next design point to minimise the expected value of $\mathcal{E}$.

# Find the minima using Bayesian optimisation

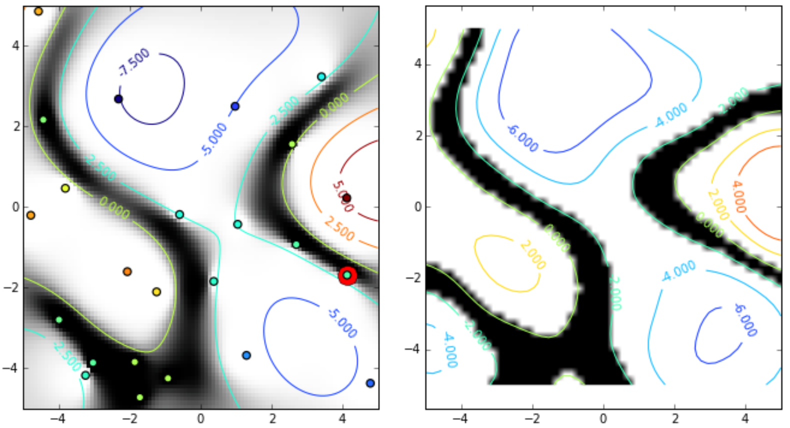
Left$=p(\theta)$, middle$=E(\theta)$, right $=\mathbb{E}(\mathcal{E}|\theta_{n+1}=\theta)$




Based on GPy and GPyOpt software.

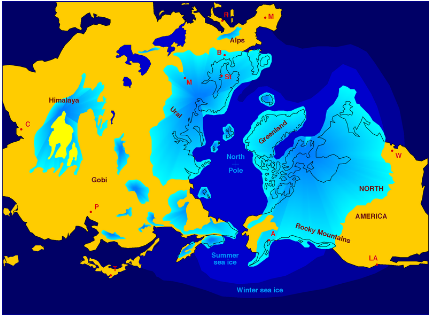
# Iteration 24

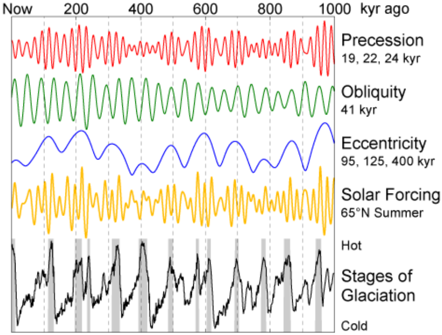
Left=estimate, right = truth



# Climate science

## What drives the glacial-interglacial cycle?




Eccentricity: orbital departure from a circle, controls duration of the seasons
Obliquity: axial tilt, controls amplitude of seasonal cycle
Precession: variation in Earth's axis of rotation, affects difference between seasons

# Model selection

What drives the glacial-interglacial cycle?

- Which aspect of the astronomical forcing is of primary importance?
- Which models best represent the cycle?

> *Most simple models of the [...] glacial cycles have at least four degrees of freedom [parameters], and some have as many as twelve. Unsurprisingly [...this is] insufficient to distinguish between the skill of the various models (Roe and Allen 1999)*

# Model selection

What drives the glacial-interglacial cycle?

- Which aspect of the astronomical forcing is of primary importance?
- Which models best represent the cycle?

> *Most simple models of the [...] glacial cycles have at least four degrees of freedom [parameters], and some have as many as twelve. Unsurprisingly [...this is] insufficient to distinguish between the skill of the various models (Roe and Allen 1999)*

Bayesian model selection revolves around the use of the Bayes factor, which are notoriously difficult to compute.

- Model selection for complex stochastic differential equations with 1000*s* data points
- Can use SMC$^2$, combined with Brownian bridge proposals, implemented on GPU

Simulation studies show we can accurately choose between competing models, and identify the correct forcing.

For real data, the age model used has an overwhelming effect on the model selection conclusions (Carson *et al.* in submission).

# Age model

Can we also quantify chronological uncertainty?

I.e., can we simultaneously date the stack, do climate reconstruction, fit the model, and choose between models?

# Age model

Can we also quantify chronological uncertainty?

Target

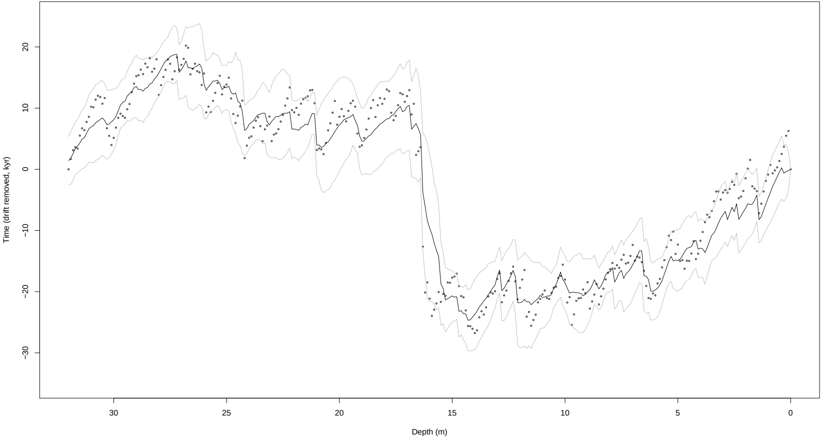
$$\pi(\theta, T_{1:N}, X_{1:N}, \mathcal{M}_k | y_{1:N})$$

where $T_{1:N}$ are the unknown times of the observations $Y_{1:N}$, $X_{1:N}$ are the climate state variables through time, $\mathcal{M}_k$ is the simulation model used, and $\theta$ is the corresponding parameter.

I.e., can we simultaneously date the stack, do climate reconstruction, fit the model, and choose between models?

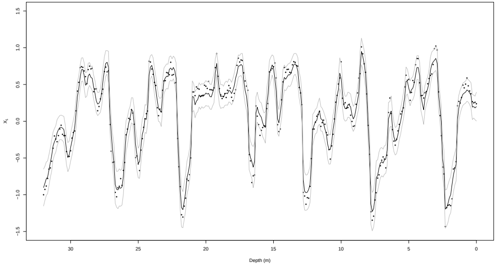
# Simulation study results - age vs depth (trend removed)

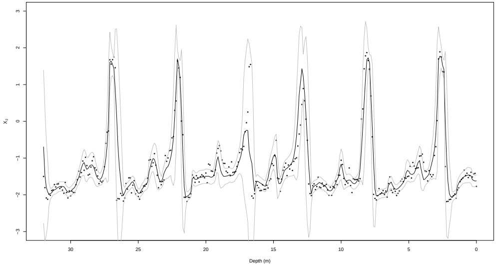
Dots = truth, black line = estimate, grey = 95% CI



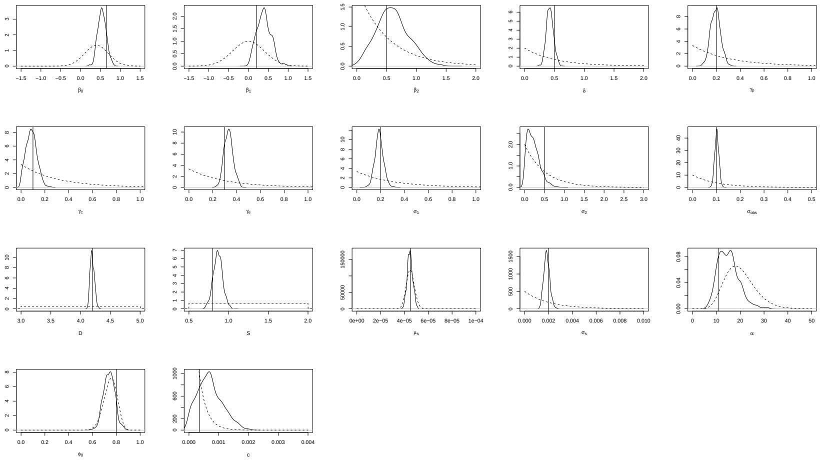
# Simulation study results - climate reconstruction

Dots = truth, black line = estimate, grey = 95% CI



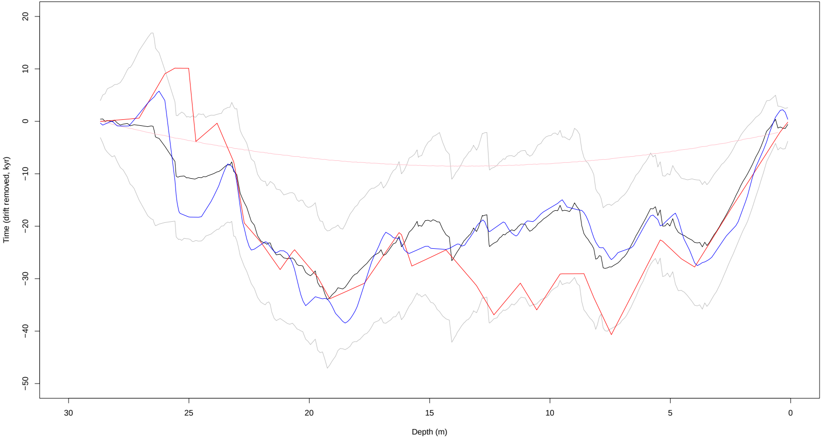
# Simulation study results - parameter estimation




Simultaneous inference of the choice between 5 models, 17 parameters, 800 ages, 1600 climate variables, using just 800 observations.

# Results for ODP846 - age vs depth (trend removed)

Black = posterior mean, grey = 95%CI, red = Huybers 2007, blue = Lisieki and Raymo 2004




Advantages: full UQ, model selection, simultaneous parameter estimation and climate reconstruction

# Model discrepancy

Consider the state space model:

$$x_{t+1} = f_\theta(x_t) + e_t, \qquad y_t = g(x_t) + \epsilon_t$$

$$e_t \sim p(\cdot), \qquad \epsilon_t \sim q(\cdot)$$

How do we correct errors in $f$ or $g$?

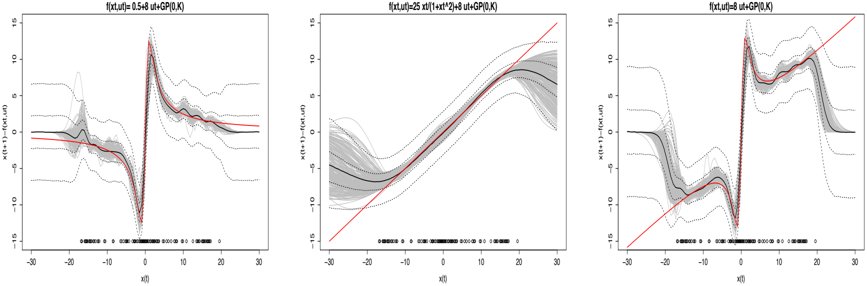
# Model discrepancy

Consider the state space model:

$$x_{t+1} = f_\theta(x_t) + e_t, \qquad y_t = g(x_t) + \epsilon_t$$

$$e_t \sim p(\cdot), \qquad \epsilon_t \sim q(\cdot)$$

How do we correct errors in $f$ or $g$?
Use a GP discrepancy model - eg, $x_{t+1} = f_\theta(x_t) + \delta(x_t) + e_t$




Technical challenge: inference using PGAS works but is expensive. A variational approach looks more promising.

## Conclusions

- David Cox: challenge for a statistician is to be involved in several fields of application and to use that to motivate theoretical contributions.
- Computational tractability is one of the key bottlenecks for much of science, either through big simulation or big data.
- Methods from machine learning have the potential to help us make large advances in statistical methodology.
  - Bayesian optimization, variational approximations, Gaussian processes, ...

# Conclusions

- David Cox: challenge for a statistician is to be involved in several fields of application and to use that to motivate theoretical contributions.
- Computational tractability is one of the key bottlenecks for much of science, either through big simulation or big data.
- Methods from machine learning have the potential to help us make large advances in statistical methodology.
  - Bayesian optimization, variational approximations, Gaussian processes, ...

Thank you for listening!