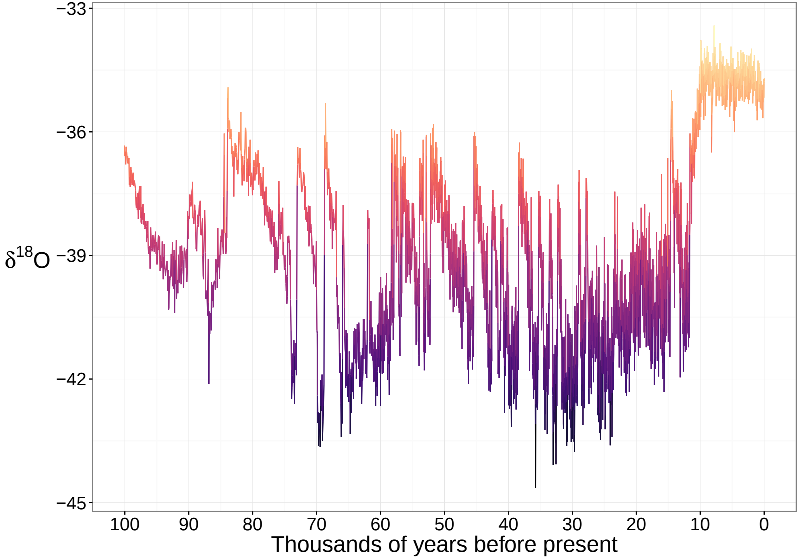
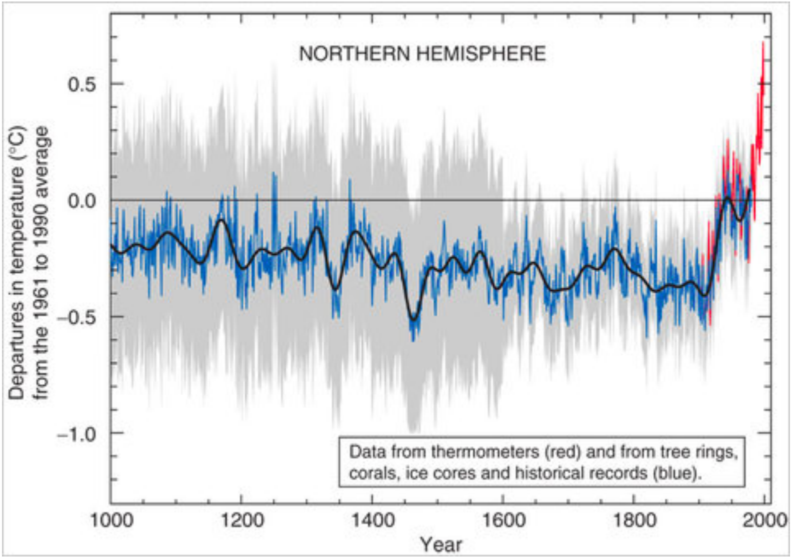
Statistical palaeoclimate reconstruction: how fast can climate change?

Andrew Parnell
andrew.parnell@ucd.ie

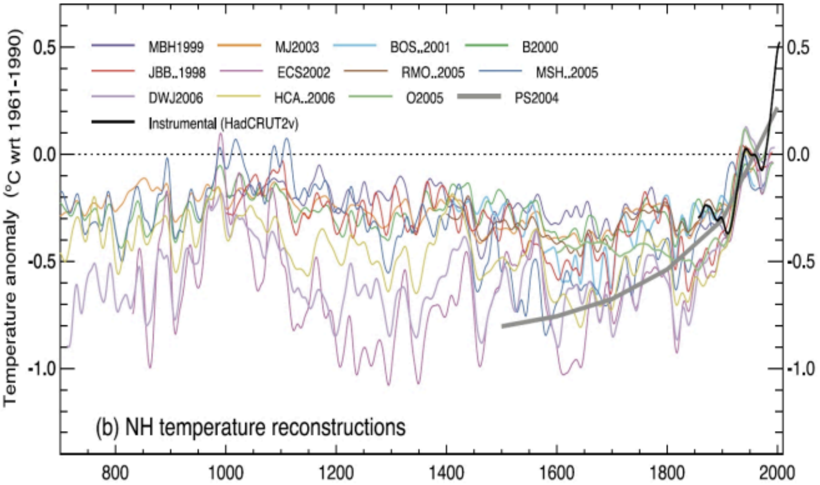
The GISP 2 ice core



The Hockey Stick

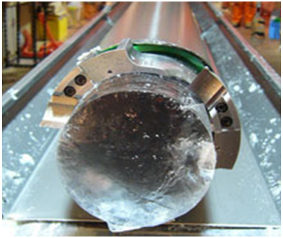


A bundle of Hockey Sticks



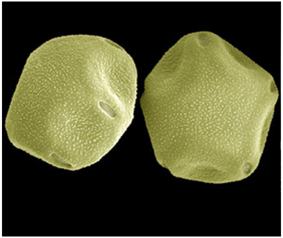
Different proxies

(a) Tree rings



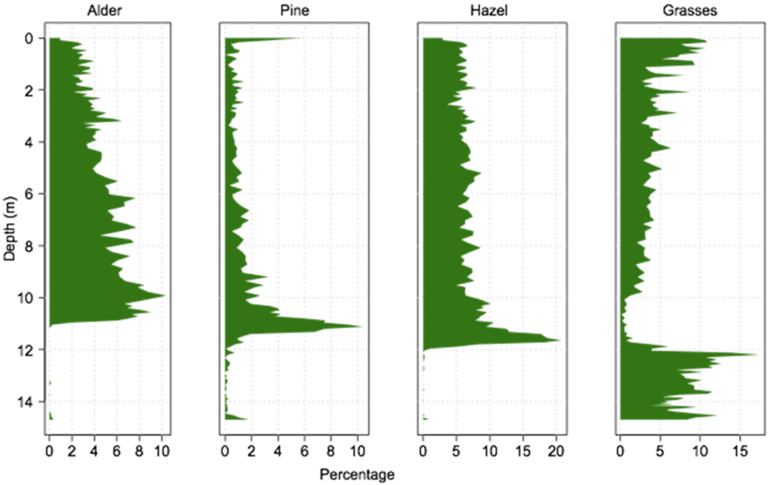
(b) An ice core

(c) A non-biting midge (chironomid)



(d) Alder pollen

Pollen depth plots



A Lago Grando di Monticchio

(By Pitichinaccio - Own work, Public Domain)

Cape May, New Jersey

(By Smallbones - Own work, CC0)

From pictures to data

Year	Climate	Proxy data			
2016	climate_{2016}	$\text{proxy}_{1,2016}$	$\text{proxy}_{2,2016}$	$\ldots$	$\text{proxy}_{p,2016}$
2015	climate_{2015}	$\text{proxy}_{1,2015}$	$\text{proxy}_{2,2015}$	$\ldots$	$\text{proxy}_{p,2015}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
1850	climate_{1850}	$\text{proxy}_{1,1850}$	$\text{proxy}_{2,1850}$	$\ldots$	$\text{proxy}_{p,1850}$
1849	$\textbf{climate}_{1849}$	$\text{proxy}_{1,1849}$	$\text{proxy}_{2,1849}$	$\ldots$	$\text{proxy}_{p,1849}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
1001	$\textbf{climate}_{1001}$	$\text{proxy}_{1,1001}$	$\text{proxy}_{2,1001}$	$\ldots$	$\text{proxy}_{p,1001}$
1000	$\textbf{climate}_{1000}$	$\text{proxy}_{1,1000}$	$\text{proxy}_{2,1000}$	$\ldots$	$\text{proxy}_{p,1000}$

A more general version

Calibration data set:

ID	Climate	Proxy data			
1	climate_1	$\text{proxy}_{1,1}$	$\text{proxy}_{2,1}$	$\ldots$	$\text{proxy}_{p,1}$
2	climate_2	$\text{proxy}_{1,2}$	$\text{proxy}_{2,2}$	$\ldots$	$\text{proxy}_{p,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
k	climate_k	$\text{proxy}_{1,k}$	$\text{proxy}_{2,k}$	$\ldots$	$\text{proxy}_{p,k}$

Fossil data set:

Year	Climate	Proxy data			
n-1	**climate_{n-1}**	$\text{proxy}_{1,n-1}$	$\text{proxy}_{2,n-1}$	$\ldots$	$\text{proxy}_{p,n-1}$
$\vdots$	**$\vdots$**	$\vdots$	$\vdots$	$\ldots$	$\vdots$
$m+1$	**climate_{m+1}**	proxy_{m+1}	proxy_{m+1}	$\ldots$	$\text{proxy}_{p,m+1}$
m	**climate_m**	$\text{proxy}_{1,m}$	$\text{proxy}_{2,m}$	$\ldots$	$\text{proxy}_{p,m}$

Some notation

Let:

- y be the ancient proxy data. Time indexed and usually multivariate
- c be ancient 'climate'. Time indexed and occasionally multivariate. Sometimes spatial too
- y^{cal} be the proxy data for the calibration period
- c^{cal} be the climate data for the calibration period

Aim is to find $c|y, y^{\text{cal}}, c^{\text{cal}}$

The regression version

Write:

$$c^{\mathsf{cal}} = f(y^{\mathsf{cal}}) + \epsilon$$

f might be a linear regression or involve some dimension reduction or variable selection.

The regression version

Write:

$$c^{\mathsf{cal}} = f(y^{\mathsf{cal}}) + \epsilon$$

f might be a linear regression or involve some dimension reduction or variable selection.

Then create:

$$\hat{c} = \hat{f}(y)$$

Problem solved!

Problems with this approach

Statistical:

- Hard to do model checking on f due to the size and nature of the calibration data
- c is often multivariate so people often pick one dimension
- The calibration period may be autocorrelated, leading to many spurious relationships
- Dimension reduction approaches will be very sensitive to the number of components chosen
- Lots of missing proxy data

Problems with this approach

Statistical:

- Hard to do model checking on f due to the size and nature of the calibration data
- c is often multivariate so people often pick one dimension
- The calibration period may be autocorrelated, leading to many spurious relationships
- Dimension reduction approaches will be very sensitive to the number of components chosen
- Lots of missing proxy data

Biological:

- The causation is the wrong way round. **Changes in climate cause changes in proxy values**
- The uncertainty in the proxies is usually substantial and not included
- The proxies might not be sensitive to northern hemisphere annual temperature, or any other chosen aspect of climate

A better way?

Instead write:

$$y^{\text{cal}} = f(c^{\text{cal}}) + \epsilon$$

f is known here as a **forward model** since it works in the causal direction. We can now include physical knowledge of how climate affects the proxies

A better way?

Instead write:

$$y^{\text{cal}} = f(c^{\text{cal}}) + \epsilon$$

f is known here as a **forward model** since it works in the causal direction. We can now include physical knowledge of how climate affects the proxies

Now **use Bayes**:

$$p(c|y, y^{\text{cal}}, c^{\text{cal}}) \propto p(y^{\text{cal}}|c^{\text{cal}})p(y|c)p(c)$$

We have the extra advantage that we can include a prior distribution $p(c)$ on the climate process

Bayesian palaeoclimate reconstruction in more detail

$$p(c, \theta, \phi | y, y^{\text{cal}}, c^{\text{cal}}) \propto p(y^{\text{cal}} | c^{\text{cal}}, \theta) p(y | c, \theta) p(c | \phi) p(\theta, \phi)$$

- $p(\theta, \phi)$ is a prior on the parameters that control the proxy/climate relationship, and climate dynamics respectively
- $p(c|\phi)$ is a prior distribution on climate dynamics. This might be a simple statistical time series model (e.g. a random walk) all the way up to a full general circulation model
- $p(y|c, \theta)$ is the forward model again, but this time applied to the missing ancient climates
- $p(y^{\text{cal}}|c^{\text{cal}}, \theta)$ is the forward model applied to the calibration data.

Why is this not the standard way people do this?

1. Building forward models is hard because you need a good calibration data set, some statistical modelling knowledge (especially with multivariate data), and some knowledge of the pollen/climate relationship
2. People want to avoid testing their models (out of sample evaluation etc)
3. Finding a good prior for climate dynamics is hard, especially if you have timing uncertainty
4. Bayes is still not common in climate science
5. Fitting this model to large calibration data sets is hard

Example: sea level rise in East Coast USA

- y^m is 18D counts of different **Foramnifera** species
- c^m is 1D - sea level
- 172 modern samples altogether
- y is also 18D made up of approx 150 layers

Fit the model:

$$p(c, \theta, \phi | y, y^{\text{cal}}, c^{\text{cal}}) \propto p(y^{\text{cal}} | c^{\text{cal}}, \theta) p(y | c, \theta) p(c | \phi) p(\theta, \phi)$$

Model details

The forward model is:

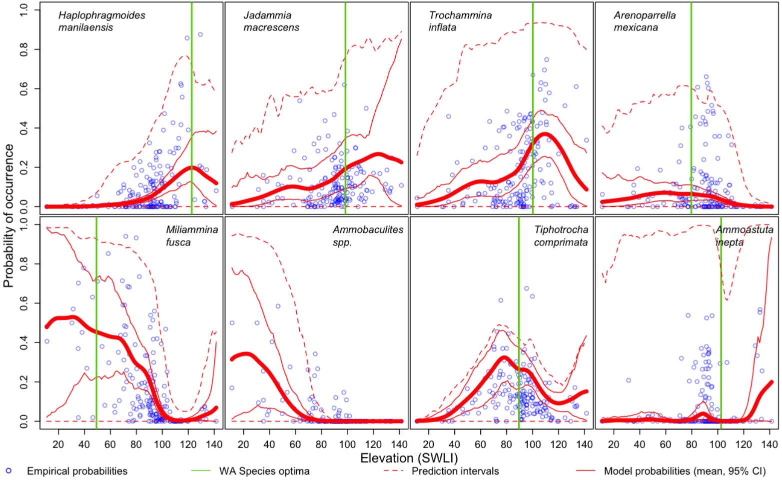
$$[y_1, \ldots, y_{28}] \sim Mult(N, \{p_1, \ldots, p_{28}\})$$

where, e.g.

$$p_l = \frac{\exp(\theta_l(c))}{\sum_j \exp(\theta_j(c))}$$

- Each θ is then given an independent P-spline with its own smoothness

Forward model output



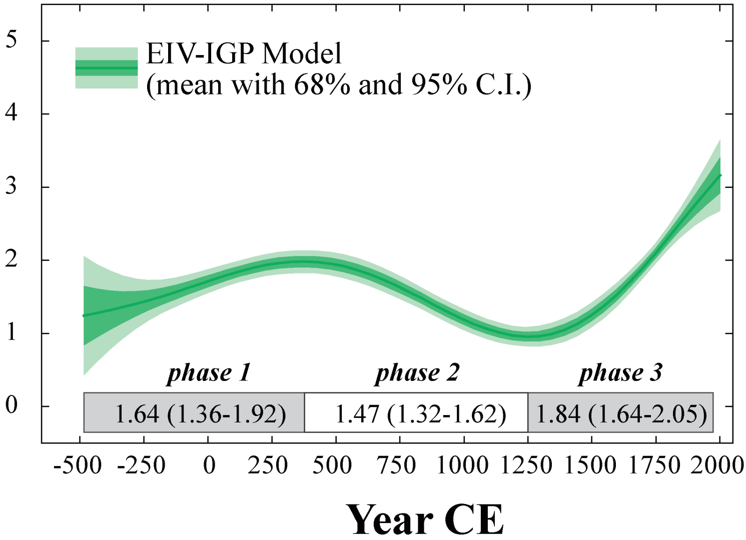
Sea level prior

- Our interest in sea level is in rates of change so we place an integrated Gaussian process prior on c over time t:

$$c(t) = \int_0^t w(u) \, \partial u$$

- We place a Gaussian process prior on w with informative priors on the mean and coavariance matrix
- The model is complicated by the presence of measurement error in time
- Our output can be either sea level c or rate w, the latter more useful

Rate of sea level rise (mm/yr) for New Jersey, USA



Example 2: multivariate climate in Italy

- This time:
  - y^m is 28D counts of different pollen types
  - c^m is 3D - two temperature and a moisture variable
  - 15000 modern samples altogether
  - y is also 28 made up of approx 900 layers
- Same model:

$$p(c, \theta, \phi | y, y^{\text{cal}}, c^{\text{cal}}) \propto p(y^{\text{cal}} | c^{\text{cal}}, \theta) p(y | c, \theta) p(c | \phi) p(\theta, \phi)$$

- Impossible to fit the above model in one go. Need approximations

Approximations

$$p(c, \theta, \phi | y, y^{\text{cal}}, c^{\text{cal}}) \propto p(y^{\text{cal}} | c^{\text{cal}}, \theta) p(y | c, \theta) p(c | \phi) p(\theta, \phi)$$

- Information on forward model parameters θ almost exclusively from modern data, so fit this separately using INLA or similar
- If the modern calibration data massively outweighs the fossil data and the prior on c is intrinsic over time, then you can show you get pretty much the full model back

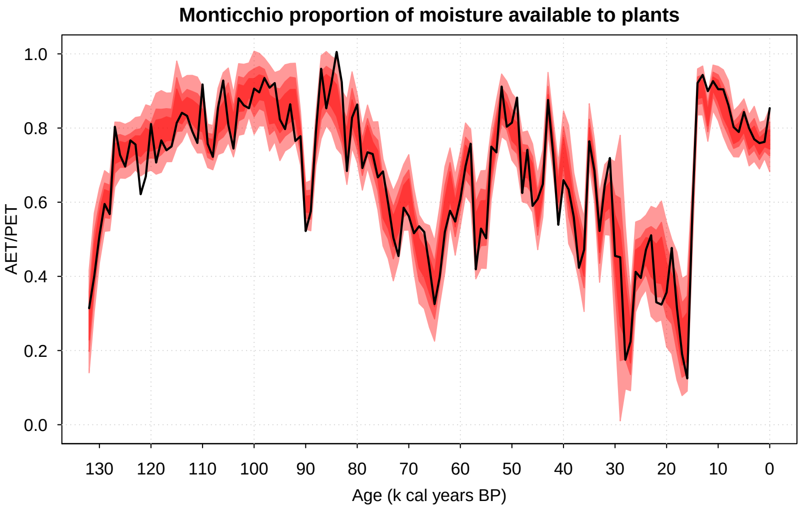
Monticchio - climate prior

- A nice intrinsic prior is a random walk
- An even nicer intrinsic prior is a Normal-Inverse Gaussian random walk

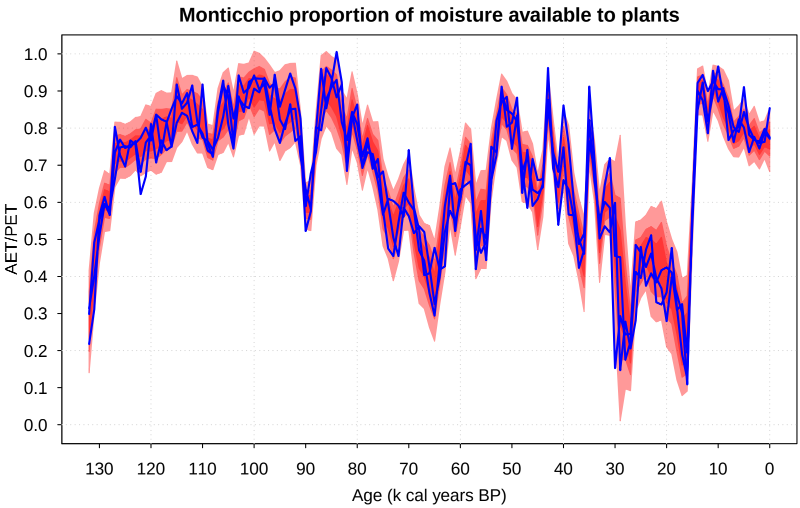
$$c(t) - c(t - \delta) \sim N(\mu, v(\delta)),\ v \sim IG(\phi_1, \phi_2)$$

- More informative priors on ϕ_1, ϕ_2
- Fit this model as a second stage using the posterior of the parameter estimates from the modern data

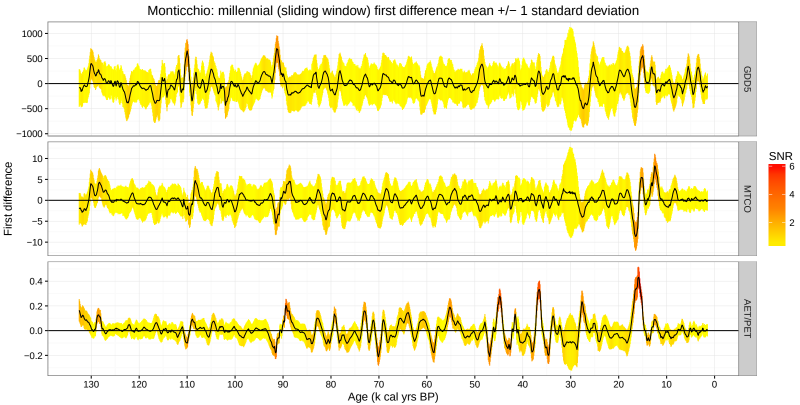
Monticchio - moisture



Example 2: Histories



Example 2: First differences - the speed of climate change



The grand challenge!

Fit a Bayesian model to:

- Reconstruct spatio-temporal palaeoclimate ...

The grand challenge!

Fit a Bayesian model to:

- Reconstruct spatio-temporal palaeoclimate ...
- ... using physical/statistical forward models for many proxies

The grand challenge!

Fit a Bayesian model to:

- Reconstruct spatio-temporal palaeoclimate ...
- ... using physical/statistical forward models for many proxies
- ... and physical/statistical models for climate dynamics

The grand challenge!

Fit a Bayesian model to:

- Reconstruct spatio-temporal palaeoclimate ...
- ... using physical/statistical forward models for many proxies
- ... and physical/statistical models for climate dynamics

The resulting output should be a large sample of spatio-temporal climate histories

Challenges 1: fitting state space models to large and complex data sets

What we really have is an **externally calibrated** state-space model in continuous time:

$$
\begin{aligned}
y^{\text{cal}}(t) &\sim f_\theta(c^{\text{cal}}(t)) \\
y(t) &\sim f_\theta(c(t)) \\
c(t) - c(t-\Delta) &\sim g_\phi(\Delta)
\end{aligned}
$$

Challenges 1: fitting state space models to large and complex data sets

What we really have is an **externally calibrated** state-space model in continuous time:

$$
\begin{aligned}
y^{\text{cal}}(t) &\sim f_\theta(c^{\text{cal}}(t)) \\
y(t) &\sim f_\theta(c(t)) \\
c(t) - c(t-\Delta) &\sim g_\phi(\Delta)
\end{aligned}
$$

- Fitting these models is hard when all the quantities are multivariate and f is a complex function
- Pseudo-marginal particle approaches seem to be the way to go for single-site models
- No obvious method yet for multi-site models. Perhaps an extension of SPDE-INLA?

Challenges 2: Incorporating mechanistic models

A more complex version:

$$\begin{aligned} y^{\text{cal}}(s,t) &\sim f_\theta(c^{\text{cal}}(s,t)) \\ y(s,t) &\sim f_\theta(c(s,t)) \\ c(s,t) &\sim g_\phi(c(\tilde{s},t^-)) \end{aligned}$$

Challenges 2: Incorporating mechanistic models

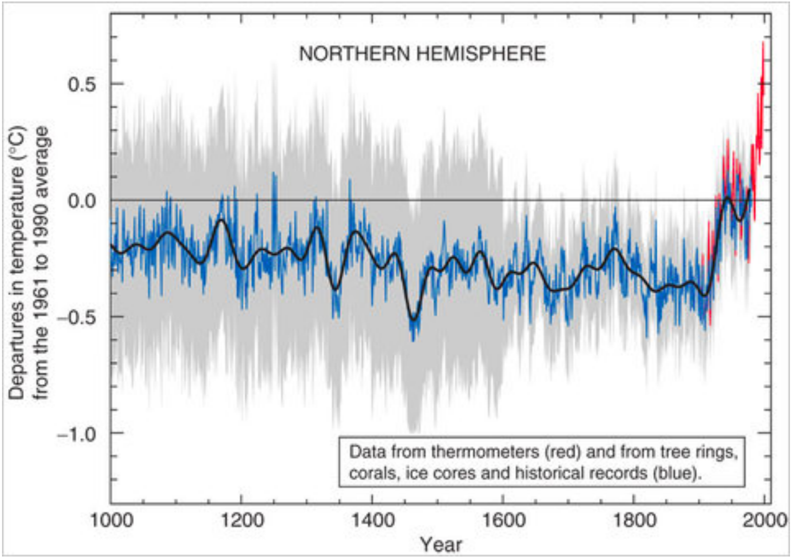
A more complex version:

$$
\begin{aligned}
y^{\text{cal}}(s,t) &\sim f_\theta(c^{\text{cal}}(s,t)) \\
y(s,t) &\sim f_\theta(c(s,t)) \\
c(s,t) &\sim g_\phi(c(\tilde{s},t^-))
\end{aligned}
$$

- The problem gets even trickier if f and g above are deterministic models
- Some quite complex deterministic models have been suggested for pollen/climate. Not many for other proxies
- Quite a few simple climate models that might work over the palaeoclimate period, e.g. Saltzman and Maasch, 1991:

$$
\begin{aligned}
dX_{(1)} &= -\left(X_{(1)} + X_{(2)} + vX_{(3)} + F(\gamma_P, \gamma_C, \gamma_E)\right) dt + \sigma_1 dW_{(1)} \\
dX_{(2)} &= \left(rX_{(2)} - pX_{(3)} - sX_{(2)}^2 - X_{(2)}^3\right) dt + \sigma_2 dW_{(2)} \\
dX_{(3)} &= -q\left(X_{(1)} + X_{(3)}\right) dt + \sigma_3 dW_{(3)}
\end{aligned}
$$

Back to the future: can we do better than this?



Summary

- A Bayesian model with an improved forward model and richer climate process for multiple sites and proxies is the ultimate research goal
- We need help with Bayesian computation for large multivariate non-linear non-Gaussian state space models
- We need help with combining deterministic/stochastic elements in forward models and climate models
- We must do better than the Hockey Stick!

References

- Cahill, N. Kemp, A.C.; Horton, B.P., and Parnell, A.C. (2015) Modeling sea-level change using errors-in-variables integrated Gaussian processes. Annals of Applied Statistics, 9(2), 547–571
- Cahill, N., Kemp, A. C., Horton, B. P., and Parnell, A. C. (2016). A Bayesian hierarchical model for reconstructing relative sea level: from raw data to rates of change. Climate of the Past, 12(2), 525–542.
- Parnell, A. C., Sweeney, J., Doan, T. K., Salter-Townshend, M., Allen, J. R., Huntley, B., and Haslett, J. (2015). Bayesian inference for palaeoclimate with time uncertainty and stochastic volatility. Journal of the Royal Statistical Society: Series C (Applied Statistics), 64(1), 115–138.