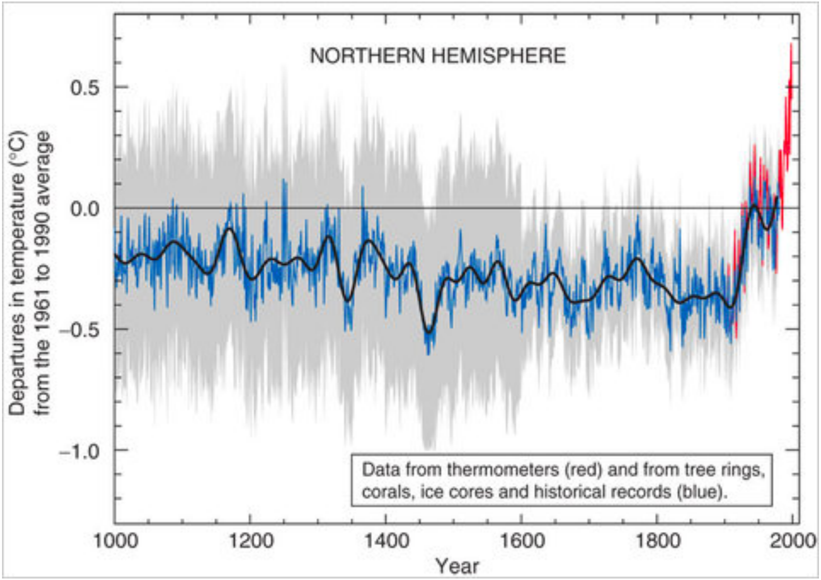
Statistical palaeoclimate reconstruction: recent results and opportunities for collaboration

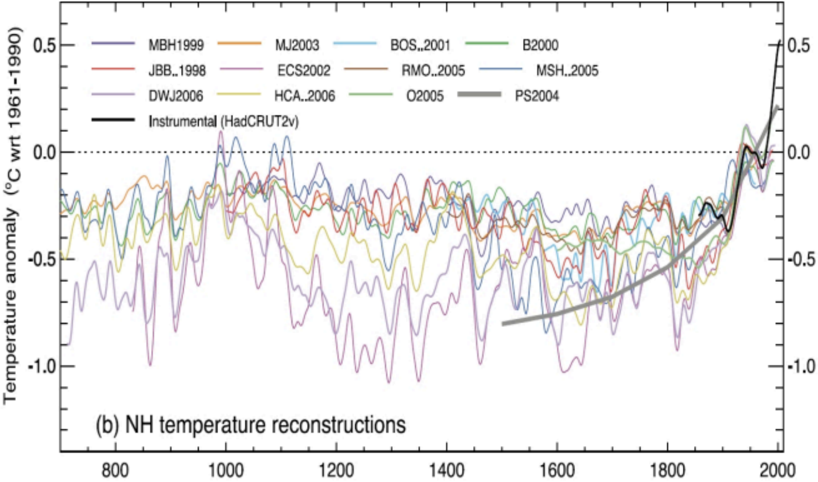
Andrew Parnell
andrew.parnell@ucd.ie

UCD School of Mathematics and Statistics

Some history...



More versions ...



How are these reconstructions created?

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	$\textbf{climate}_{1000}$	$\text{proxy}_{1,1000}$	$\text{proxy}_{2,1000}$	$\ldots$	$\text{proxy}_{p,1000}$

Or more generally...

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	$\text{proxy}_{1,2}$	$\text{proxy}_{2,2}$	$\ldots$	$\text{proxy}_{p,2}$
⋮	⋮	⋮	⋮	$\ldots$	⋮
k	climate_k	$\text{proxy}_{1,k}$	$\text{proxy}_{2,k}$	$\ldots$	$\text{proxy}_{p,k}$

Palaeoclimate:

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

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

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

The regression version

Write:

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

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
- The calibration period is 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
- The calibration period is 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 Bayesian 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 include physical knowledge of how climate affects the proxies

A better Bayesian 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 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

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

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.

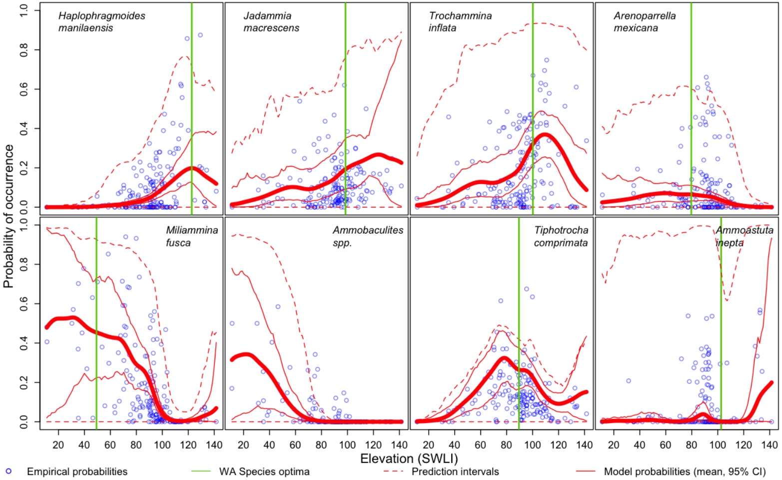
Why is this not the standard way people do this?

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

Example: sea level rise in East Coast USA

- **Foramnifera** (or forams) live in the tidal range along coastal marshes
- There are lots of different species, and they all like slightly different parts of the tidal range
- If you take a sediment core on the marsh you can count lots of fossilised forams (which can also be dated) and produce a history of sea level height at that site
- We also take a number of surface samples from the local region to build up a calibration data set of which forams like which aspect of the tidal range

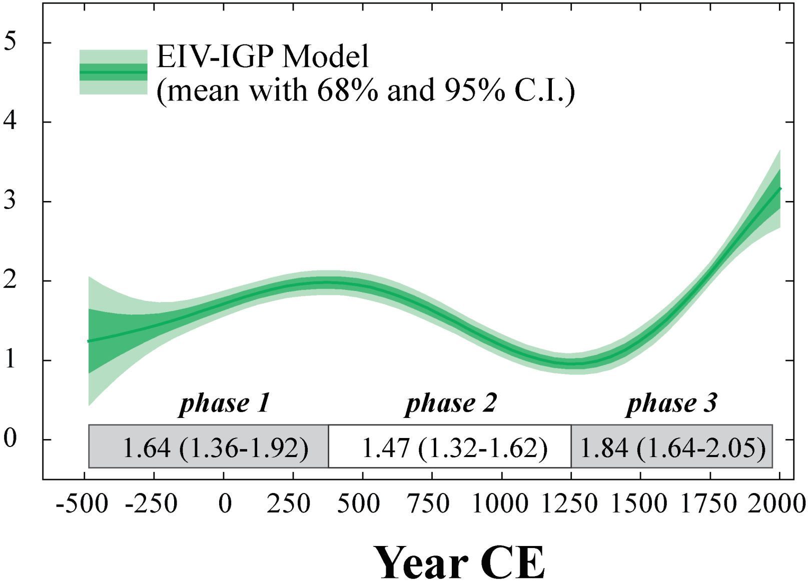
The forward model



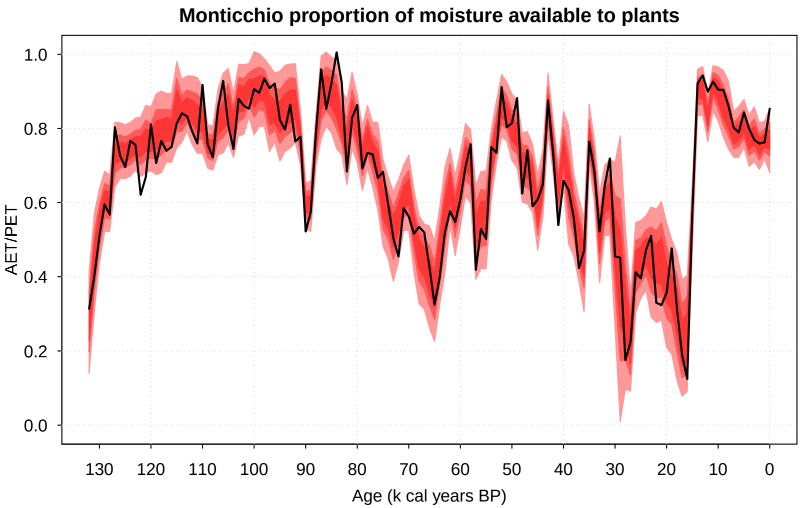
Model description

- Our forward model for the forams uses multinomial counts and P-splines
- We have a second proxy (called $\delta^{13}C$) that gives further information on the position in the tidal frame at that depth in the core
- Our prior on climate dynamics (here height of sea level over time) uses a fancy Gaussian process

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



Example 2: multivariate climate in Italy



Other examples:

- Parnell, A. C., Sweeney, J., Doan, T. K., Salter-Townshend, M., Allen, J. R. M., Huntley, B., & 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.
- Tolwinski-Ward, S. E., Tingley, M. P., Evans, M. N., Hughes, M. K., & Nychka, D. W. (2014). Probabilistic reconstructions of local temperature and soil moisture from tree-ring data with potentially time-varying climatic response. Climate Dynamics, 44(3-4), 791–806.
- Holmström, L., Ilvonen, L., Seppä, H., & Veski, S. (2015). A Bayesian spatiotemporal model for reconstructing climate from multiple pollen records. The Annals of Applied Statistics, 9(3), 1194–1225.

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 am **externally calibrated** state-space model in continuous time:

$$
\begin{aligned}
y^{\text{cal}}(t) &= f(c^{\text{cal}}(t)) + \epsilon^{\text{cal}} \\
y(t) &= f(c(t)) + \epsilon
\end{aligned}
$$

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

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

$$
\begin{aligned}
y^{\text{cal}}(t) &= f(c^{\text{cal}}(t)) + \epsilon^{\text{cal}} \\
y(t) &= f(c(t)) + \epsilon \\
c(t) &= c(t - \Delta) + \gamma
\end{aligned}
$$

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

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

$$
\begin{aligned}
y^{\text{cal}}(t) &= f(c^{\text{cal}}(t)) + \epsilon^{\text{cal}} \\
y(t) &= f(c(t)) + \epsilon \\
c(t) &= c(t - \Delta) + \gamma
\end{aligned}
$$

- 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 SPDE-INLA?

Challenges 2: Incorporating mechanistic models

A more complex version (ignoring external calibration):

$$\begin{aligned} y(t) &= f(c(t)) \\ c(t) &= g(c(t_-) \end{aligned}$$

Challenges 2: Incorporating mechanistic models

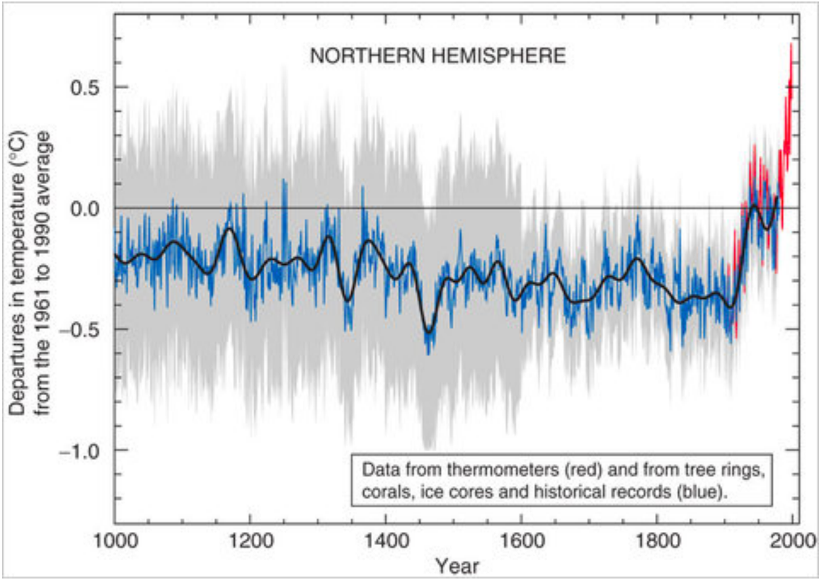
A more complex version (ignoring external calibration):

$$
\begin{aligned}
y(t) &= f(c(t)) \\
c(t) &= g(c(t_-)
\end{aligned}
$$

- The problem gets 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. SM91:

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

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



Summary

- A Bayesian version model with good forward models which produces climate histories seems like the best way to go for this work
- 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 can do better than the Hockey Stick!