# Hurrah for Proxy Data!

Jonathan Rougier

Dept. Mathematical Sciences, University of Durham, UK

# Abstract

Acceptable inferences for future climate need to be constrained by a range of different types of data, taken not just from the climate state vector, but also from physical and biological processes that are affected by climate (“proxy data”). I explain why this is, what types of proxy data might be used, and, looking to the future, how they should be incorporated.

# Computer Experiments

- *Computer Experiment*: (Probabilistic) inference about a system using a computer simulator:

$$\texttt{Model} + \texttt{Treatment} + \texttt{Solver} = \texttt{Simulator}$$

# Computer Experiments

- *Computer Experiment*: (Probabilistic) inference about a system using a computer simulator:

$$\texttt{Model} + \texttt{Treatment} + \texttt{Solver} = \texttt{Simulator}$$

- Sources of uncertainty:
  - Uncertain model parameters
  - Simulator inadequacy (a.k.a. "structural error")
  - Measurement error on system data used for calibration

# Computer Experiments

- *Computer Experiment*: (Probabilistic) inference about a system using a computer simulator:

$$\texttt{Model} + \texttt{Treatment} + \texttt{Solver} = \texttt{Simulator}$$

- Sources of uncertainty:
  - Uncertain model parameters
  - Simulator inadequacy (a.k.a. “structural error”)
  - Measurement error on system data used for calibration
- Statistical insights can help with
  - Specifying prior uncertainties
  - Understanding the inferential calculation
  - Choosing informative evaluations
  - Handling slow simulators / large parameter spaces

# Climate Sensitivity

- **Question:** What is the probability that global mean climate will be $2^{\circ}$C warmer in 2100?

# Climate Sensitivity

- **Question:** What is the probability that global mean climate will be $2^\circ$C warmer in 2100?
- *Climate sensitivity* is the change in steady-state global mean temperature following a doubling of atmospheric $CO_2$. We denote it as $\lambda$. It is a key input into climate change impact assessment, and into the development of response strategies (mitigate, adapt).

# Climate Sensitivity

- **Question:** What is the probability that global mean climate will be $2^\circ$C warmer in 2100?
- *Climate sensitivity* is the change in steady-state global mean temperature following a doubling of atmospheric $CO_2$. We denote it as $\lambda$. It is a key input into climate change impact assessment, and into the development of response strategies (mitigate, adapt).
- We ought to be able to compute it using a computer simulator of climate, $g(\cdot)$ say. The problem is that we are not sure about the correct parameterisation of the simulator. In a nutshell, we need to estimate

$$F_\lambda(\ell) \triangleq \text{Pr}(\lambda \leq \ell) \qquad \text{where} \qquad \lambda \triangleq g_\lambda(x^*)$$

and $x^*$ is uncertain, with probability distribution function $F_{x^*}$.

# Climate Sensitivity

- **Question:** What is the probability that global mean climate will be $2^\circ$C warmer in 2100?
- *Climate sensitivity* is the change in steady-state global mean temperature following a doubling of atmospheric $CO_2$. We denote it as $\lambda$. It is a key input into climate change impact assessment, and into the development of response strategies (mitigate, adapt).
- We ought to be able to compute it using a computer simulator of climate, $g(\cdot)$ say. The problem is that we are not sure about the correct parameterisation of the simulator. In a nutshell, we need to estimate

$$F_\lambda(\ell) \triangleq \Pr(\lambda \leq \ell) \qquad \text{where} \qquad \lambda \triangleq g_\lambda(x^*)$$

  and $x^*$ is uncertain, with probability distribution function $F_{x^*}$.
- To help us, we will *calibrate* our climate simulator using observations on actual climate.

# The prior predictive distribution

- $F_\lambda$ is known as the *prior predictive distribution*. Formally we write it as

$$F_\lambda(\ell) = \int_x \mathbf{1}\big(g_\lambda(x) \leq \ell\big)\ dF_{x^*}(x)$$

where $\mathbf{1}(\cdot)$ is the indicator function. This operation sums the probability content of the region of $x^*$ for which the simulator gives a climate sensitivity value less than $\ell$.

# The prior predictive distribution

- $F_\lambda$ is known as the *prior predictive distribution*. Formally we write it as

$$F_\lambda(\ell) = \int_x \mathbf{1}\big(g_\lambda(x) \le \ell\big)\; dF_{x^*}(x)$$

where $\mathbf{1}(\cdot)$ is the indicator function. This operation sums the probability content of the region of $x^*$ for which the simulator gives a climate sensitivity value less than $\ell$.

- The simplest way to estimate $F_\lambda$ is by *Monte Carlo integration*:

$$\hat{F}_\lambda^{(n)}(\ell) \triangleq n^{-1} \sum_{i=1}^{n} \mathbf{1}\big(g_\lambda(X_i) \le \ell\big)$$

where $X_1, \ldots, X_n$ are independently sampled from $F_{x^*}$.

# The prior predictive distribution

- $F_\lambda$ is known as the *prior predictive distribution*. Formally we write it as

$$F_\lambda(\ell) = \int_x \mathbf{1}\big(g_\lambda(x) \leq \ell\big)\ dF_{x^*}(x)$$

where $\mathbf{1}(\cdot)$ is the indicator function. This operation sums the probability content of the region of $x^*$ for which the simulator gives a climate sensitivity value less than $\ell$.

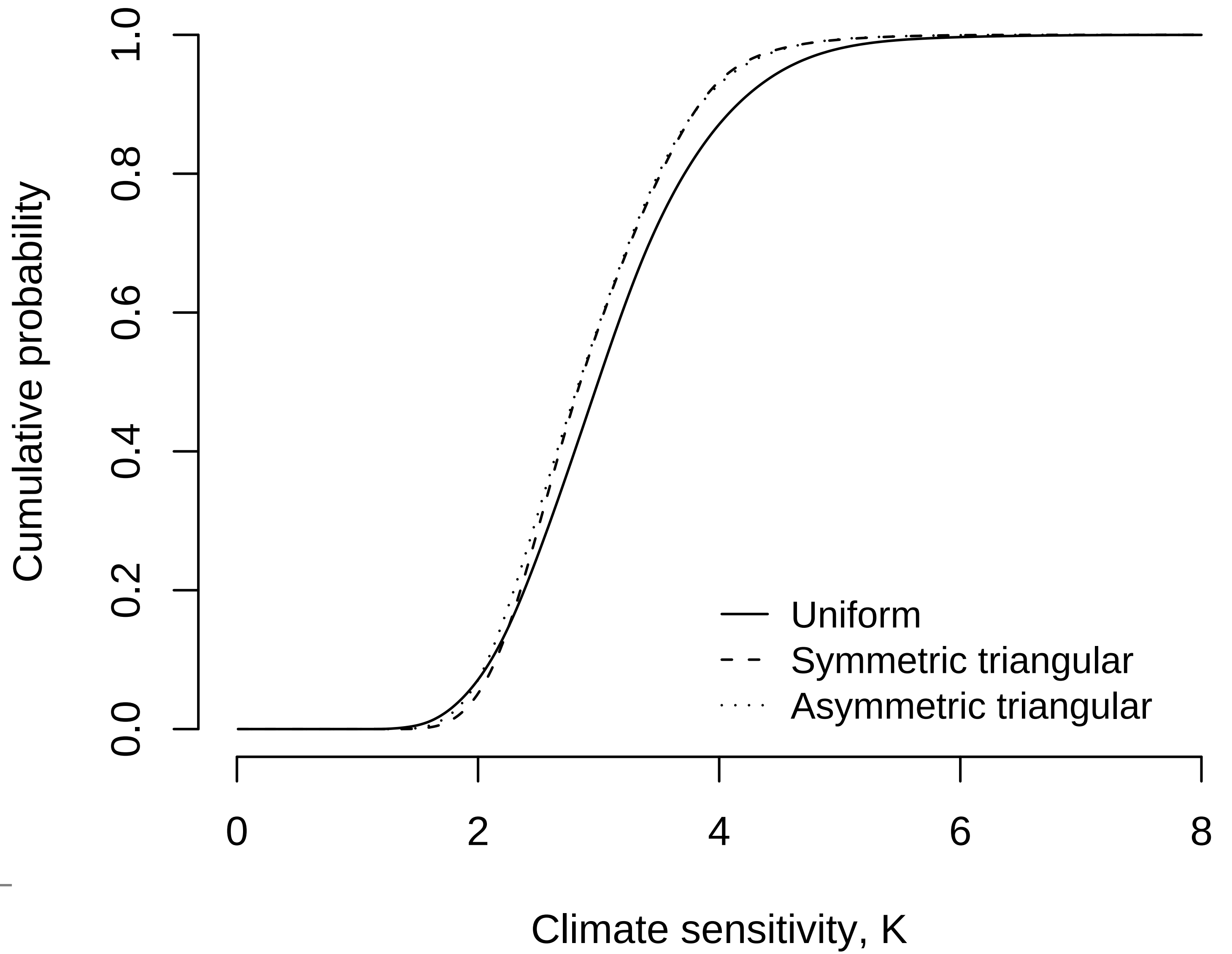
- The simplest way to estimate $F_\lambda$ is by *Monte Carlo integration*:

$$\hat{F}_\lambda^{(n)}(\ell) \triangleq n^{-1} \sum_{i=1}^{n} \mathbf{1}\big(g_\lambda(X_i) \leq \ell\big)$$

where $X_1, \ldots, X_n$ are independently sampled from $F_{x^*}$.

- By the *Strong Law of Large Numbers*, we have $\lim_{n\to\infty} \hat{F}_\lambda^{(n)}(\ell) \to F_\lambda(\ell)$. There are lots of ways we might improve our estimate of $F_\lambda$, for example *importance sampling with variance reduction techniques*.

# An example, following Murphy et al (2004)



# The posterior predictive distribution

- This distribution is much more interesting! It incorporates information about $x^*$ from actual climate data.

# The posterior predictive distribution

- This distribution is much more interesting! It incorporates information about $x^*$ from actual climate data.
- Applying *Bayes's theorem*,

$$F_{\lambda|z}(\ell) \triangleq \text{Pr}(\lambda \leq \ell \mid z = \tilde{z})$$
$$= c \int_x \mathbf{1}\big(g_\lambda(x) \leq \ell\big) \; \text{Lik}_{\tilde{z}}(x) \; dF_{x^*}(x)$$

where $c \triangleq \text{Pr}(z = \tilde{z})^{-1}$ and $\text{Lik}_{\tilde{z}}(x) \triangleq \text{Pr}(z = \tilde{z} \mid x^* = x)$.

Here $z$ denotes the actual climate data, and $\tilde{z}$ their observed values.

# The posterior predictive distribution

- This distribution is much more interesting! It incorporates information about $x^*$ from actual climate data.

- Applying *Bayes's theorem*,

$$F_{\lambda \mid z}(\ell) \triangleq \text{Pr}(\lambda \leq \ell \mid z = \tilde{z})$$
$$= c \int_x \mathbf{1}\big(g_\lambda(x) \leq \ell\big) \; \text{Lik}_{\tilde{z}}(x) \; dF_{x^*}(x)$$

where $c \triangleq \text{Pr}(z = \tilde{z})^{-1}$ and $\text{Lik}_{\tilde{z}}(x) \triangleq \text{Pr}(z = \tilde{z} \mid x^* = x)$.

Here $z$ denotes the actual climate data, and $\tilde{z}$ their observed values.

- In order to specify the *likelihood function* $\text{Lik}_{\tilde{z}}(\cdot)$, we need a statistical model linking $x^*$ and $z$; for example

$$z = g_z(x^*) + \epsilon + e$$

where $x^*$, $\epsilon$ and $e$ are mutually independent, and $(\epsilon, e) \sim$ Gaussian.

# The posterior PD (cont)

- Under our assumptions we have

$$\mathsf{Lik}_{\tilde{z}}(x) = \phi\big(\tilde{z}\ ;\ g_z(x), \Sigma^{\epsilon} + \Sigma^{e}\big)$$

where $\phi(\cdot\ ;\ \cdot, \cdot)$ is a gaussian probability density function with given mean vector and variance matrix (we must specify $\Sigma^{\epsilon}$ and $\Sigma^{e}$).

# The posterior PD (cont)

- Under our assumptions we have

$$\mathsf{Lik}_{\tilde{z}}(x) = \phi\big(\tilde{z}\ ;\, g_z(x), \Sigma^{\epsilon} + \Sigma^{e}\big)$$

where $\phi(\cdot\ ;\ \cdot, \cdot)$ is a gaussian probability density function with given mean vector and variance matrix (we must specify $\Sigma^{\epsilon}$ and $\Sigma^{e}$).

- Now we can estimate $F_{\lambda|z}$ using the Monte Carlo approach:

$$\hat{F}^{(n)}_{\lambda|z}(\ell) \triangleq \sum_{i=1}^{n} w_i\, \mathbf{1}\big(g_\lambda(x) \leq \ell\big)$$

where $w_i \propto \mathsf{Lik}_{\tilde{z}}(X_i)$ and $\sum_{i=1}^{n} w_i = 1$, and $X_1, \dots, X_n$ are sampled as before.

# The posterior PD (cont)

- Under our assumptions we have

$$\mathsf{Lik}_{\tilde{z}}(x) = \phi\big(\tilde{z}\,;\, g_z(x), \Sigma^\epsilon + \Sigma^e\big)$$

  where $\phi(\cdot\,;\,\cdot,\cdot)$ is a gaussian probability density function with given mean vector and variance matrix (we must specify $\Sigma^\epsilon$ and $\Sigma^e$).

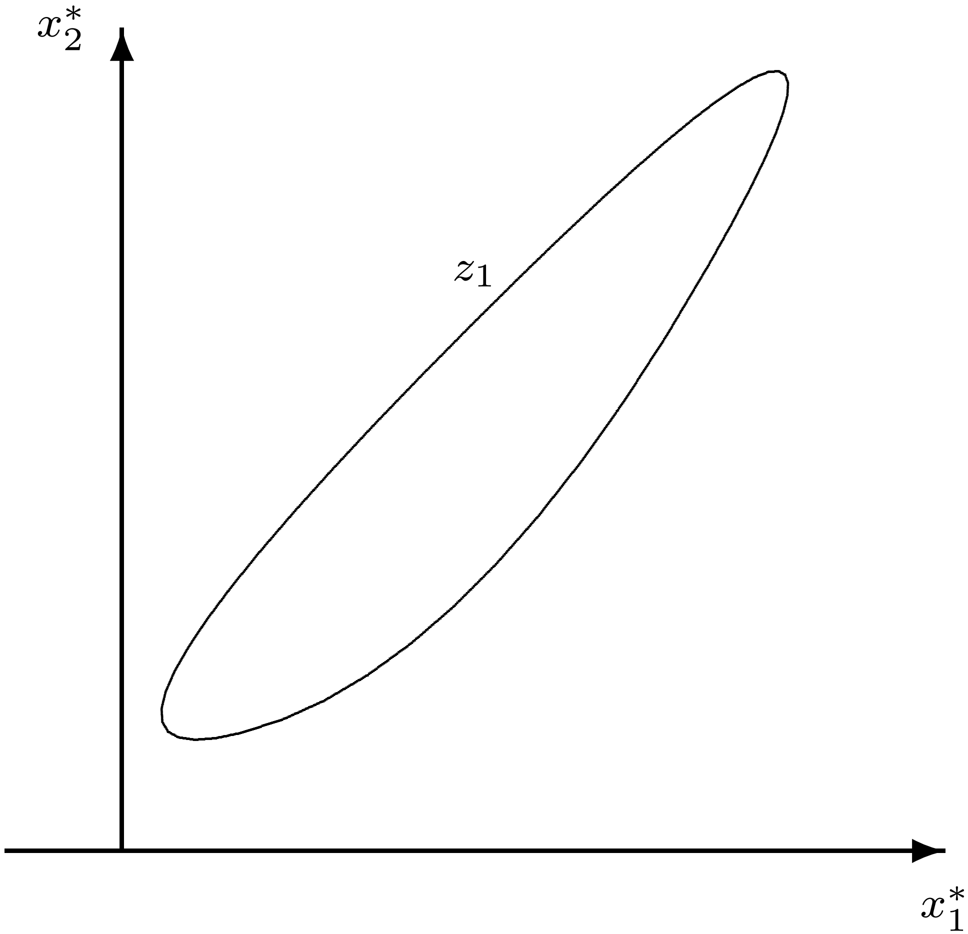
- Now we can estimate $F_{\lambda|z}$ using the Monte Carlo approach:

$$\hat{F}^{(n)}_{\lambda|z}(\ell) \triangleq \sum_{i=1}^{n} w_i\, \mathbf{1}\big(g_\lambda(x) \leq \ell\big)$$

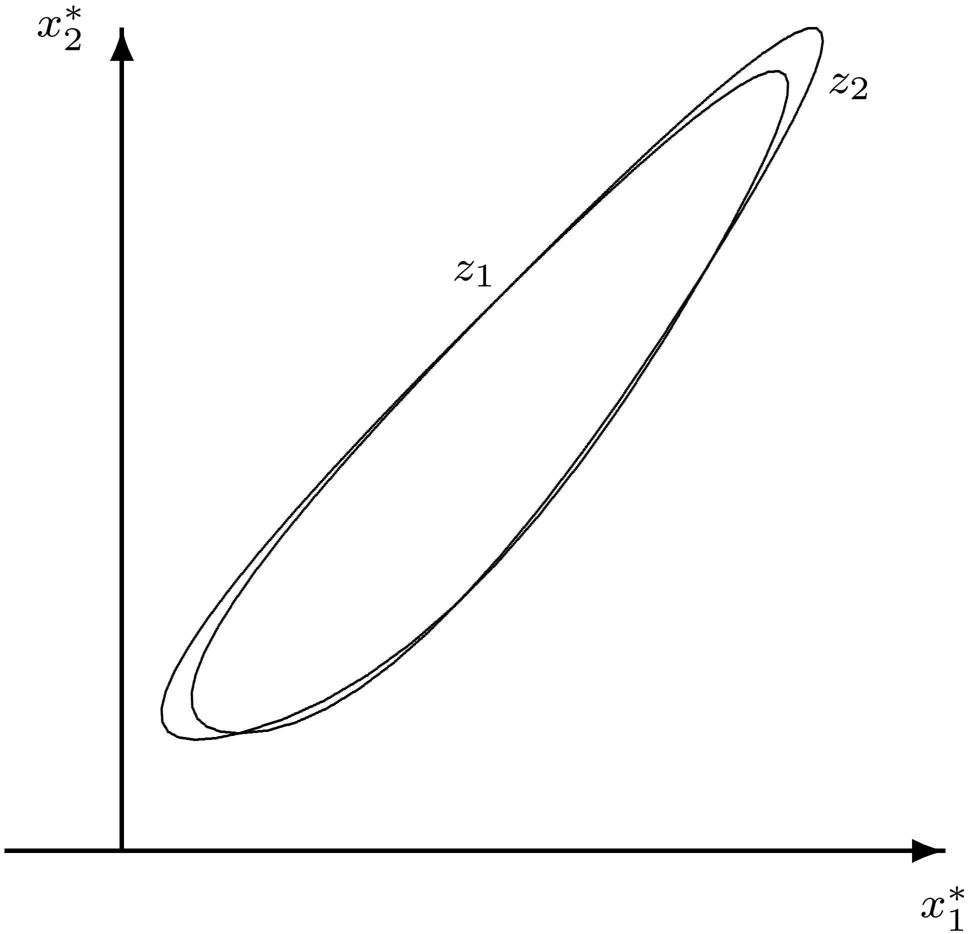
  where $w_i \propto \mathsf{Lik}_{\tilde{z}}(X_i)$ and $\sum_{i=1}^n w_i = 1$, and $X_1, \dots, X_n$ are sampled as before.

- For the prior predictive distribution we had $w_i \propto 1$. *The effect of the data $z = \tilde{z}$ is to down-weight the contribution of candidate values for $x^*$ for which the simulator is not able to replicate the actual data $\tilde{z}$.*

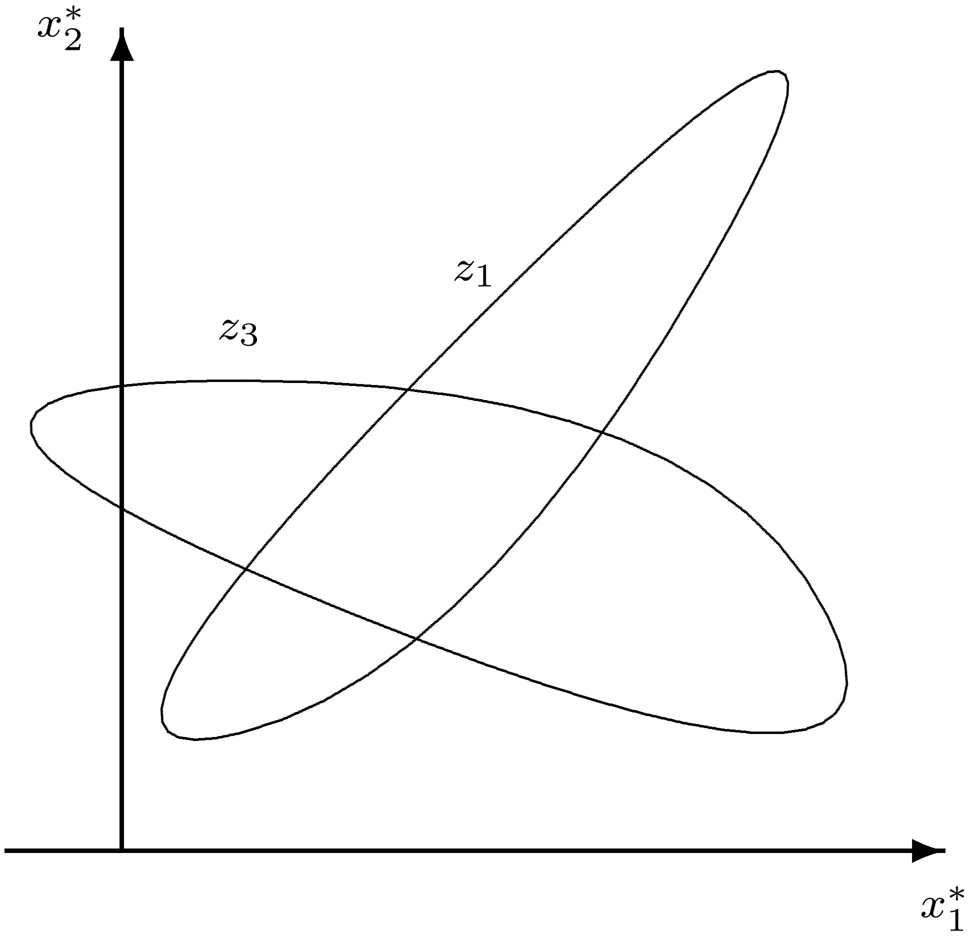
# The role of the data $z$



# The role of the data $z$



# The role of the data $z$



# Types of data

- In the Bayesian paradigm, the role of the prior distribution $F_{x^*}$ is diminished as the amount of data becomes large, and the likelihood function becomes 'concentrated'.

# Types of data

- In the Bayesian paradigm, the role of the prior distribution $F_{x^*}$ is diminished as the amount of data becomes large, and the likelihood function becomes 'concentrated'.
- The standard data comprises observations on the *climate state vector*, i.e. measurements that can be mapped directly to the output of the climate simulator: temperature, pressure, salinity, humidity, velocity, clouds *etc.*. These measurements are plentiful, but they do not constrain $x^*$ enough.

# Types of data

- In the Bayesian paradigm, the role of the prior distribution $F_{x^*}$ is diminished as the amount of data becomes large, and the likelihood function becomes ‘concentrated’.
- The standard data comprises observations on the *climate state vector*, i.e. measurements that can be mapped directly to the output of the climate simulator: temperature, pressure, salinity, humidity, velocity, clouds *etc.*. These measurements are plentiful, but they do not constrain $x^*$ enough.
- *Proxy data* are observations made on processes that are affected by climate. They are quite different from the climate state vector, and quite different from each other.

# Types of data

- In the Bayesian paradigm, the role of the prior distribution $F_{x^*}$ is diminished as the amount of data becomes large, and the likelihood function becomes ‘concentrated’.
- The standard data comprises observations on the *climate state vector*, i.e. measurements that can be mapped directly to the output of the climate simulator: temperature, pressure, salinity, humidity, velocity, clouds *etc.*. These measurements are plentiful, but they do not constrain $x^*$ enough.
- *Proxy data* are observations made on processes that are affected by climate. They are quite different from the climate state vector, and quite different from each other.
- My favourite sources of proxy data:
  - Fossilised tree-rings (dendrochronology)
  - Sedimentary record of temperature-sensitive organisms
  - Oceanic water oxygen-isotope ratio
  - Composition of atmospheric bubbles in ice-cores
  - Geological evidence from strata

# Including proxy data in the calibration

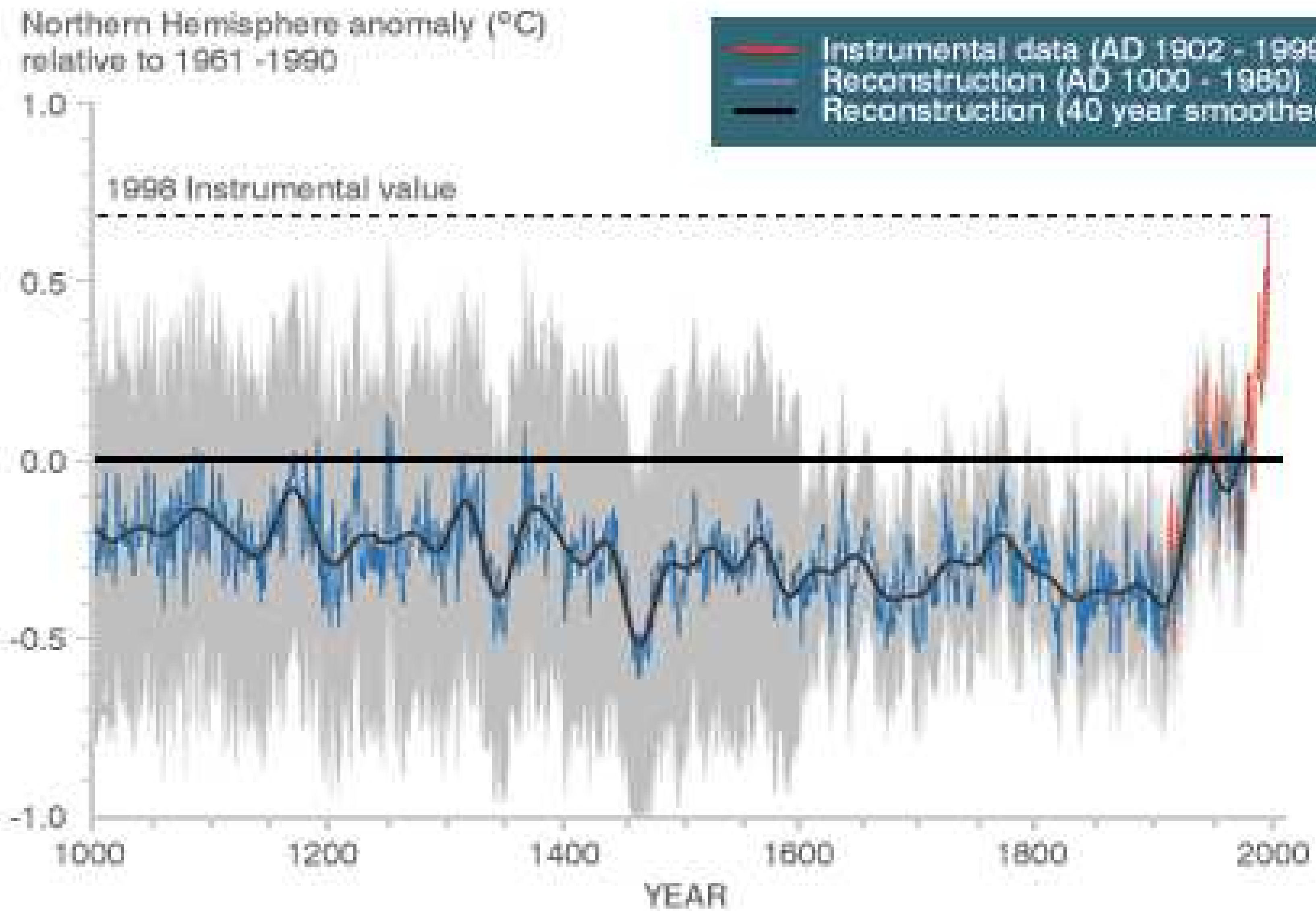
- If $z$ is our proxy data, and $y$ is the true value of the climate state vector, then typically we can construct a 'forward' mapping of the form

$$(x^*, y) \mapsto z$$

  where $x^*$ contains important climate forcing variables like atmospheric $CO_2$, or spatio-temporal descriptors such as land-use and forestation.

- *The wrong way*: Try and turn the proxy data into measurements on the climate state vector. This will not work well because the 'forward' mapping is not invertible in the form $z \to y$. But that has not stopped the climate scientists!

# Including proxy data in the calibration



# Including proxy data in the calibration

- If $z$ is our proxy data, and $y$ is the true value of the climate state vector, then typically we can construct a 'forward' mapping of the form

$$(x^*, y) \rightarrow z$$

where $x^*$ contains important climate forcing variables like atmospheric $CO_2$, or spatio-temporal descriptors such as land-use and forestation.

- *The wrong way*: Try and turn the proxy data into measurements on the climate state vector. This will not work well because the 'forward' mapping is not invertible in the form $z \rightarrow y$. But that has not stopped the climate scientists!

# Including proxy data in the calibration

- If $z$ is our proxy data, and $y$ is the true value of the climate state vector, then typically we can construct a 'forward' mapping of the form

$$(x^*, y) \to z$$

  where $x^*$ contains important climate forcing variables like atmospheric $CO_2$, or spatio-temporal descriptors such as land-use and forestation.

- *The wrong way*: Try and turn the proxy data into measurements on the climate state vector. This will not work well because the 'forward' mapping is not invertible in the form $z \to y$. But that has not stopped the climate scientists!

- *The right way*: Add the forward mapping to the simulator, and include the proxy data in the simulator output. The rationale: *The Bayesian approach is going to solve the inverse problem anyway, as long as we can write down the forward model.*

# Conclusion

- If you are doing a computer experiment and you want to make probabilistic inferences, there is a large body of literature to help
- Different computer experiments have different problems; with climate prediction, one problem is that the data are not sufficiently differentiated for a useful calibration
- Any data that are affected by the system being simulated can be used for calibration, if we can construct a ‘forward model’
- Proxy data in climate, including biological data, could be disproportionately useful in calibrating climate simulators.