1b. Scientific V&V

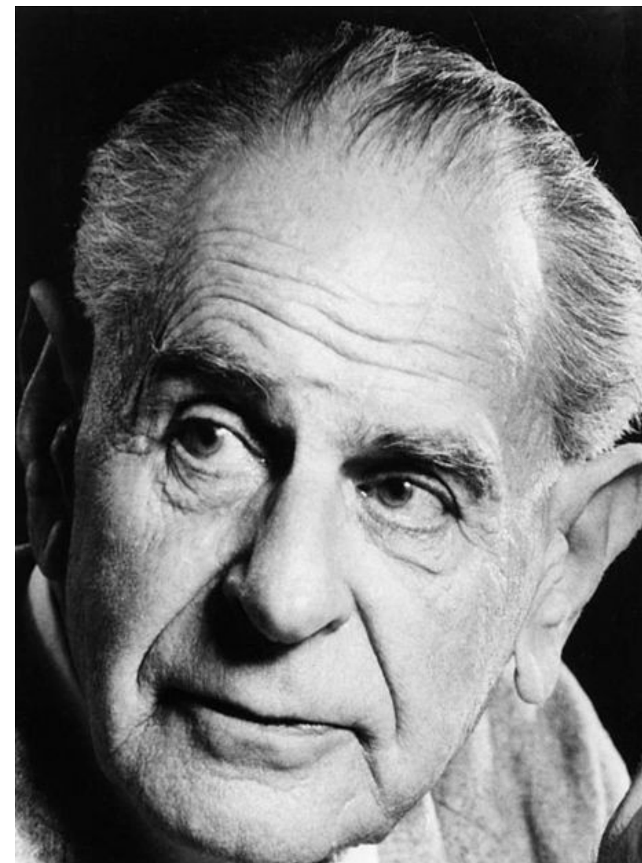
Falsification
Literature research
Scientific replication

Rasmus E. Benestad
Winter School in eScience
Geilo January 20-25, 2013

Falsification

Science: **hypotheses cannot be verified.**
Hypotheses can be **falsified.**

Aim: to look for consistency
Falsification.



Karl Popper
From Wikipedia, the free encyclopedia

Critical search for inconsistencies

Bring in all available relevant information

What have others found?

Literature research – **independent** studies.

- Peer reviewed – some quality control (?)
- **Understand** the analysis and science
- **Trace** information through threads of references
- **Published papers must be replicable too**
- Danger in falling into dogma - sloppiness

Propagation of error through citation

Tempting to cite papers not read, or not check that the paper actually supports claim (not caught by review).

Example: Tropical Cyclones (TC) and an oft-cited statement: *area of warm ocean does not affect the cyclone frequency*.

Benestad, R. E. 'An Explanation for the Lack of Trend in the Hurricane Frequency'. arXiv:physics/0603195 (March 23, 2006). http://arxiv.org/abs/physics/0603195:

"...the **thermodynamic technique cited by Henderson-Sellers *et al.* (1998) is tailored for the intensity of TCs rather than their frequency**. The statement about the relationship between the warm area and cyclogenesis [generation of cyclones] is re-examined ... Henderson-Sellers *et al.* (1998) do not provide convincing evidence for why the cyclogenseis should not be sensitive to warm pool area".

The responsibility of a scientist

- Read and understand the analysis.
- Trace key references to source.
- Repeat the work – replicate
  - Lab experiments
  - Numerical analysis/simulations
- Differences – how to resolve?
  - More details: sciencequestions

Scientific replication

“many published results are impossible to reproduce”.

Replications should be replicable.

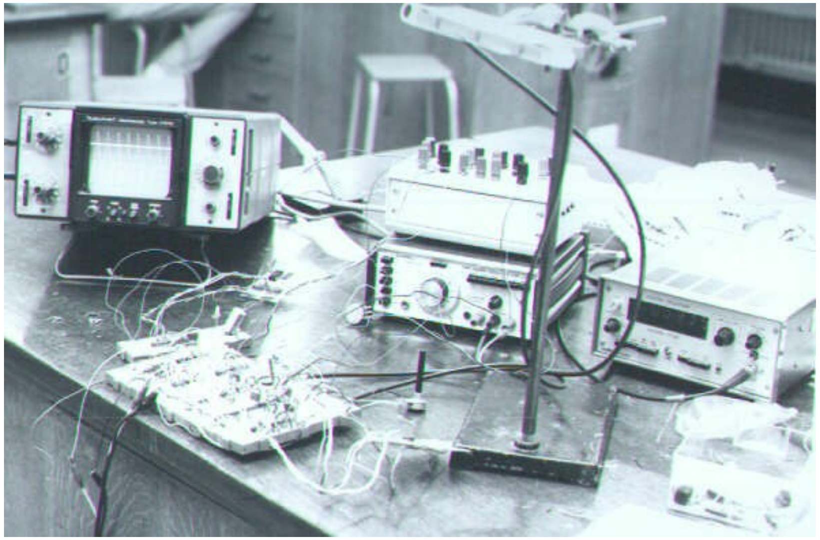
Science is about universal truths – the general features must be reproducible, otherwise

- **non-robust**
- **weak signal (insignificant)**
- **not objective**

Types of scientific replication

Lab demonstrations – important role, however, not in the scope of these lectures.

Here: **Computer-based replication.**



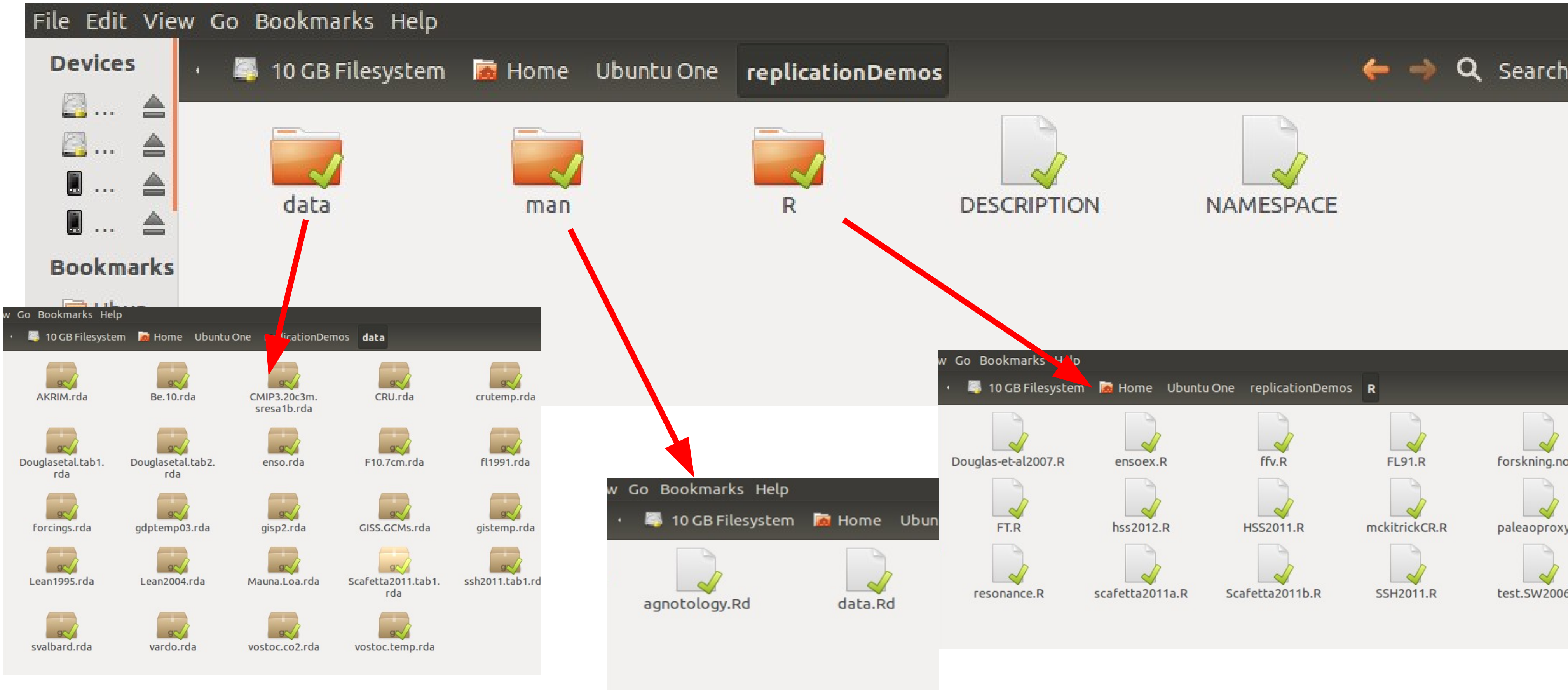
Replication and numerical analysis

- E.g. R-packages & R-scripts.
- Important considerations for quality & traceability
  - Signature and in-line comments
- Tests to verify previous results.
- Test the tests...
  - Design code to test the key functions
  - Sample data – hypothetical cases

R-packages

- Ordered information – version control.
- Well structured documentation.
  - Browser-based, hyperlinked, PDF, searchable.
- Open source code.
- Data.
- Demonstrations & examples.
- Based on long experience (S++, S, ...)

R-packages



- **Pebesma**, E., D. Nüst, and R. Bivand **(2012)**, The R Software Environment in Reproducible Geoscientific Research, ***Eos***, Vol. 93, No. 16, 17, p. 163-164.

Example: 'replicationDemos'

- R-package addressing '**agnotology**':
- Open-source, open data, replication & testing
- Number of different case studies, taken from the scientific literature.
- Tables digitally copied from the PDF-versions of the paper.
- Data – with URL attribute for identifying sources.
- Traceability
- **How do we arrive at the results?**

'Cooking' recipes

```
File  Edit  Options  Buffers  Tools  Imenu-S  ESS  Help

Douglass2007 <- function() {

  df2m <- function(X) {
# Convert the data.frame into a matrix:
    #print("df2m:")
    v <- names(X)[-(1:2)]
    d <- dim(X)
    #print(d)
    d[2] <- length(v)
    M <- matrix(rep(NA,d[1]*d[2]),d[1],d[2])
    for (i in 1:d[2])
      eval(parse(text=paste("M[,i]<-X$",v[i],sep="")))
    colnames(M) <- substr(v,2,nchar(v))
    rownames(M) <- X$runs
    #print("M:"); print(M)
    invisible(M)
  }


  p.hydrostatic <- function (h, p0 = 1000, Temp = 288, g = 9.81,
                             k = 1.38e-23, M = 0.027/6.022e+23)
{
    p <- p0 * exp(-(M * g * h)/(k * Temp))
    p
}

  cat("Reproduction of results in Fig 1. of Douglas et al. (2007)")
  cat("'A comparison of tropical temperature trends with model predictions'")
  cat("INTERNATIONAL JOURNAL OF CLIMATOLOGY")
  cat("Int. J. Climatol. (2007)")
  cat("Published online in Wiley InterScience")
  cat("(www.interscience.wiley.com) DOI: 10.1002/joc.1651")
  cat("")
  cat("Based on Tables I & II in the paper. The values have been")
  cat("copied from the on-line PDF through acroreader.")
  cat("(the negative sign of the values had to be set to '-')")

  data("Douglasetal.tab1",envir=environment())
  #load("Debunking/data/Douglasetal.tab1.rda")
  data("Douglasetal.tab2",envir=environment())
  #load("Debunking/data/Douglasetal.tab2.rda")
  X1 <- df2m(Douglasetal.tab1)/1000
  lev1 <- attr(Douglasetal.tab1,'levels')
  X2 <- df2m(Douglasetal.tab2)/1000
  dim(X2) <- c(22,13)
  #print(class(X2))
  lev2 <- attr(Douglasetal.tab2,'levels')
  plot(range(100,1000),c(-0.5,1.5),type="n",
--:---  Douglas-et-al2007.R   Top L47    (ESS[S] [none] Rox)---------------------
tool-bar kill-buffer
```

Scripts facilitate exact replication

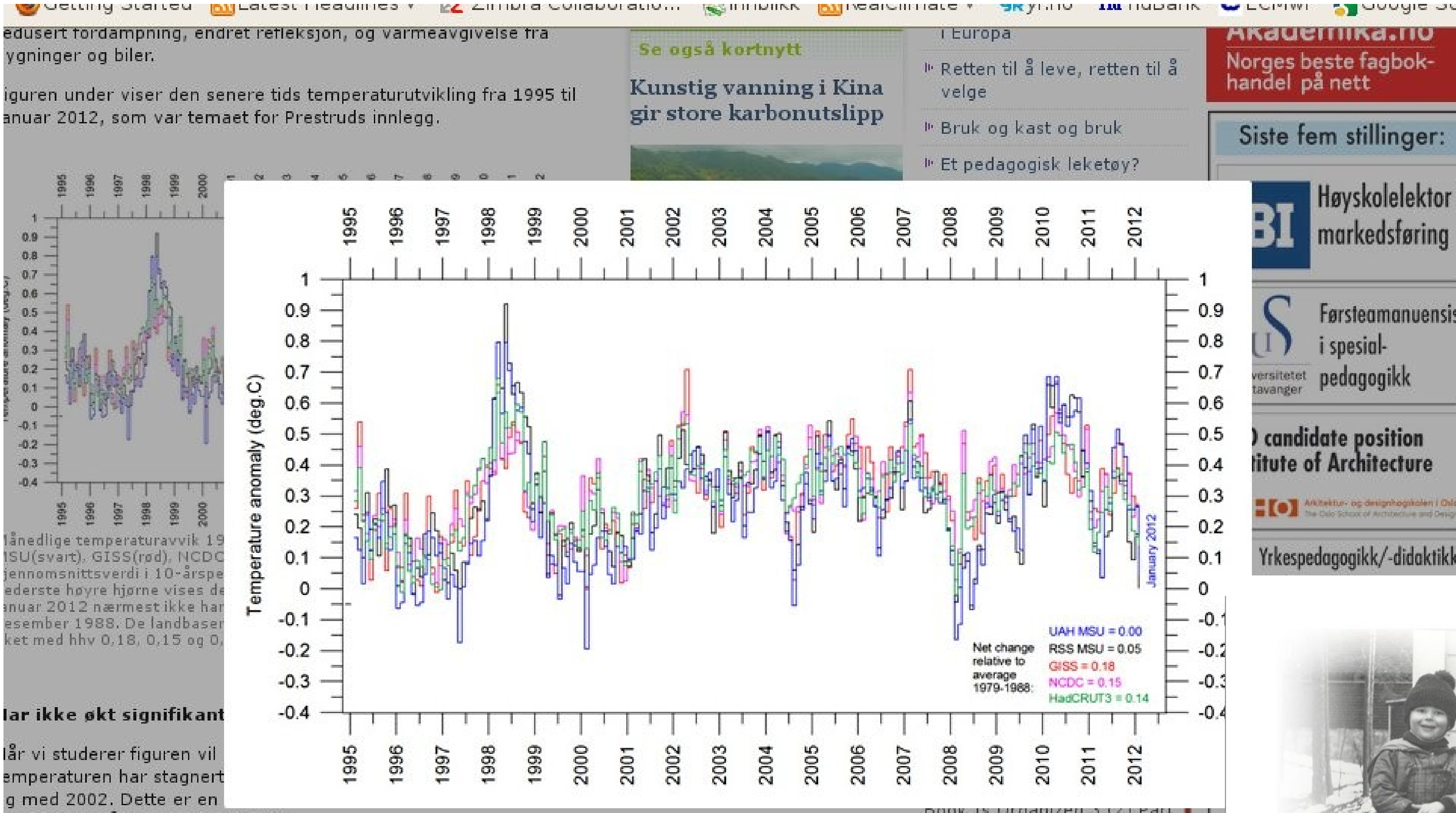
```
File  Edit  Options  Buffers  Tools  Rd  Help

\name{svalbard}
\alias{AKRIM}
\alias{CRU}
\alias{crutemp}
\alias{F10.7cm}
\alias{forcings}
\alias{GISS.GCMs}
\alias{gistemp}
\alias{Lean1995}
\alias{Lean2004}
\alias{svalbard}
\alias{ssh2011.tab1}
\alias{vardo}
\alias{Douglasetal.tab1}
\alias{Douglasetal.tab2}
\alias{Mauna.Loa}
\alias{gdptemp03}
\alias{gisp2}
\alias{enso}
\alias{vostoc.co2}
\alias{vostoc.temp}
\alias{Be.10}
\alias{CMIP3.20c3m.sresa1b}
\alias{Scafetta2011.tab1}
\alias{fl1991}
\title{Data for demonstrations of replication and testing.}
\description{
Various data sets used in the demonstrations. Several of these are
'standard' data sets (CRU, Lean2004, AKRIM, crutemp, F10.7cm, forcings,
gistemp, Lean1995, GISP2, Mauna.Loa). Some are from tables in papers
(tab1, Douglasetal.tab1, Douglasetal.tab2,Scafetta2011.tab1).

The tables were copied digitally from the PDF-version in acroreader
(copy text) and then saves as ASCII-files, read in R, and then re-saved as
rda-files. The negative signs ('-') had to be set to '-' since the ASCII
code for the signs in the tables did not correspond to the ASCII code
used by R. Once these minor issues were fixed, these should be exact
reproductions of the tables in the papers.

\code{ssh2011.tab1} is the data from Table 1 in Solheim et al. (2011)
\code{Douglasetal.tab1} and \code{Douglasetal.tab1} are from Douglas et
al.

The other data sets have been taken from the same sources as stated in
the papers. The URL from where these were obtained are given in the data
attributes (e.g. type \code{names(attributes(gisp2))}).

By copying the numbers in published tables, and providing these together
--:---  data.Rd        Top L1      (Rd Fill)-----------------------------------
Rd mode version 0.9-1
```

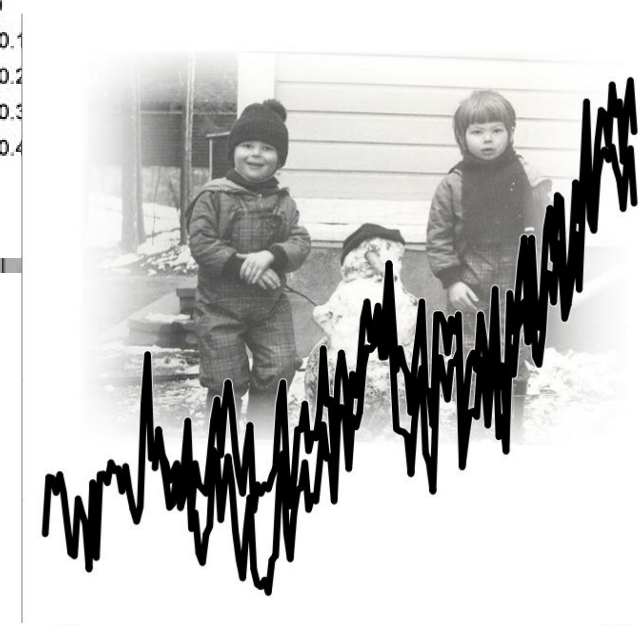
Case studies:

- Examples from climate research.
- Real-life controversies
- Claims:
  - “The global warming has stopped”
  - “The climate is driven by Jupiter, Saturn and the moon”
  - “Climate models don't account for the observed role from Jupiter”

“The global warming has stopped”

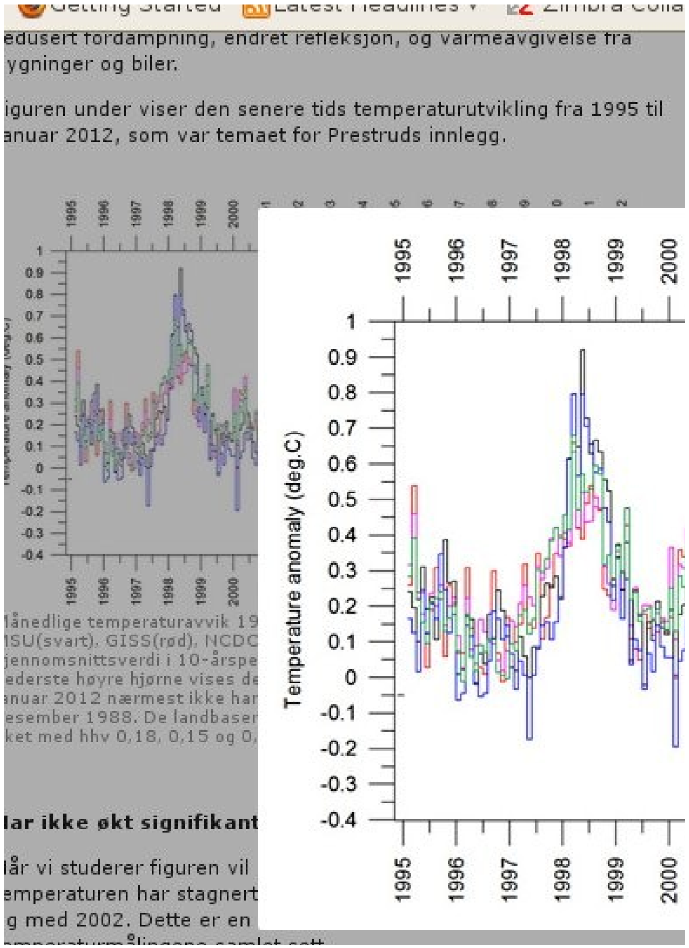
Case 1: A global warming hiatus?

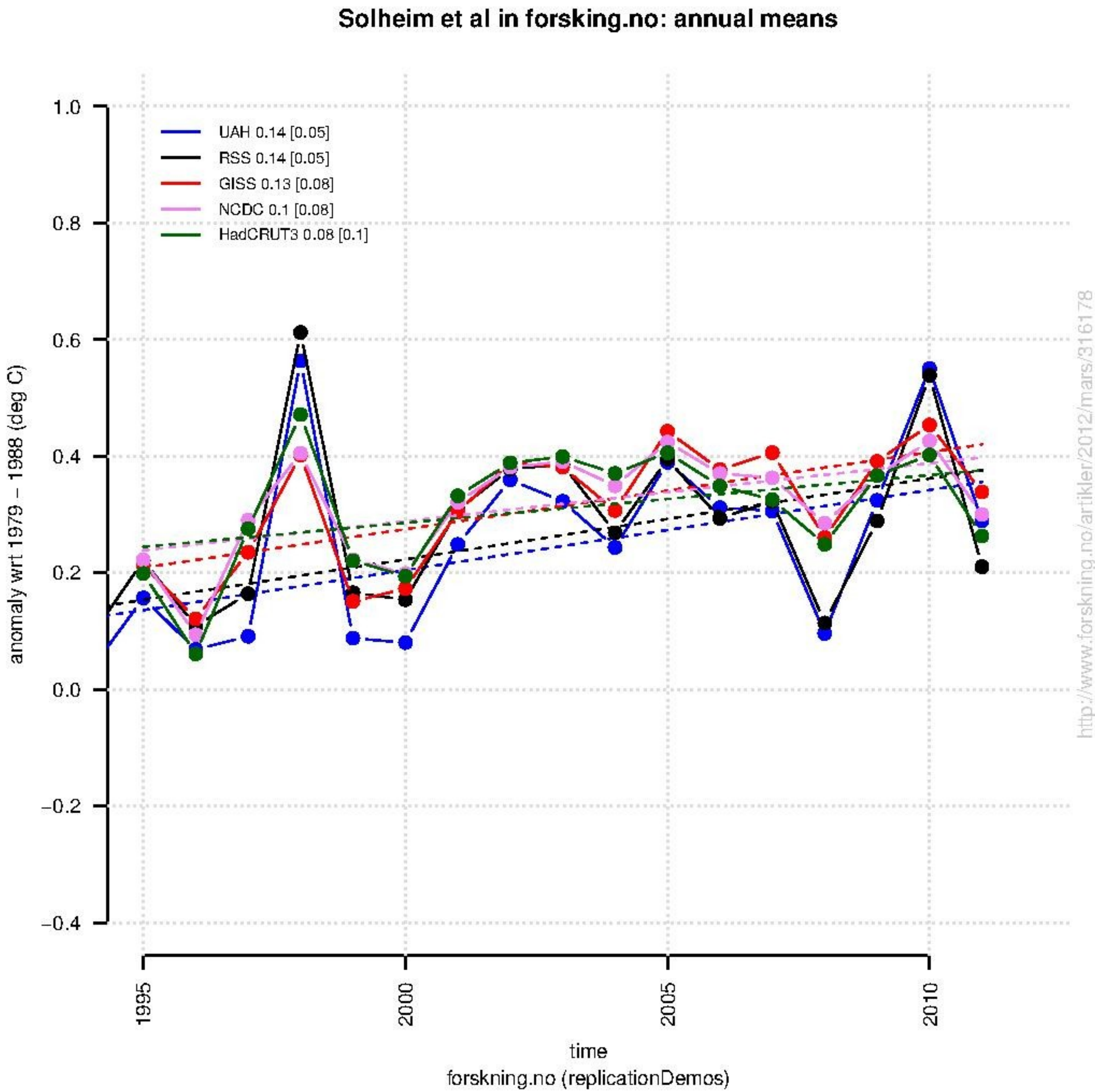




“The global warming has stopped”

Test - regression





Same data

Different emphasis

Case 2:Replication of prediction

Humlum et al. (2011), *Glob. Planet. Change*:

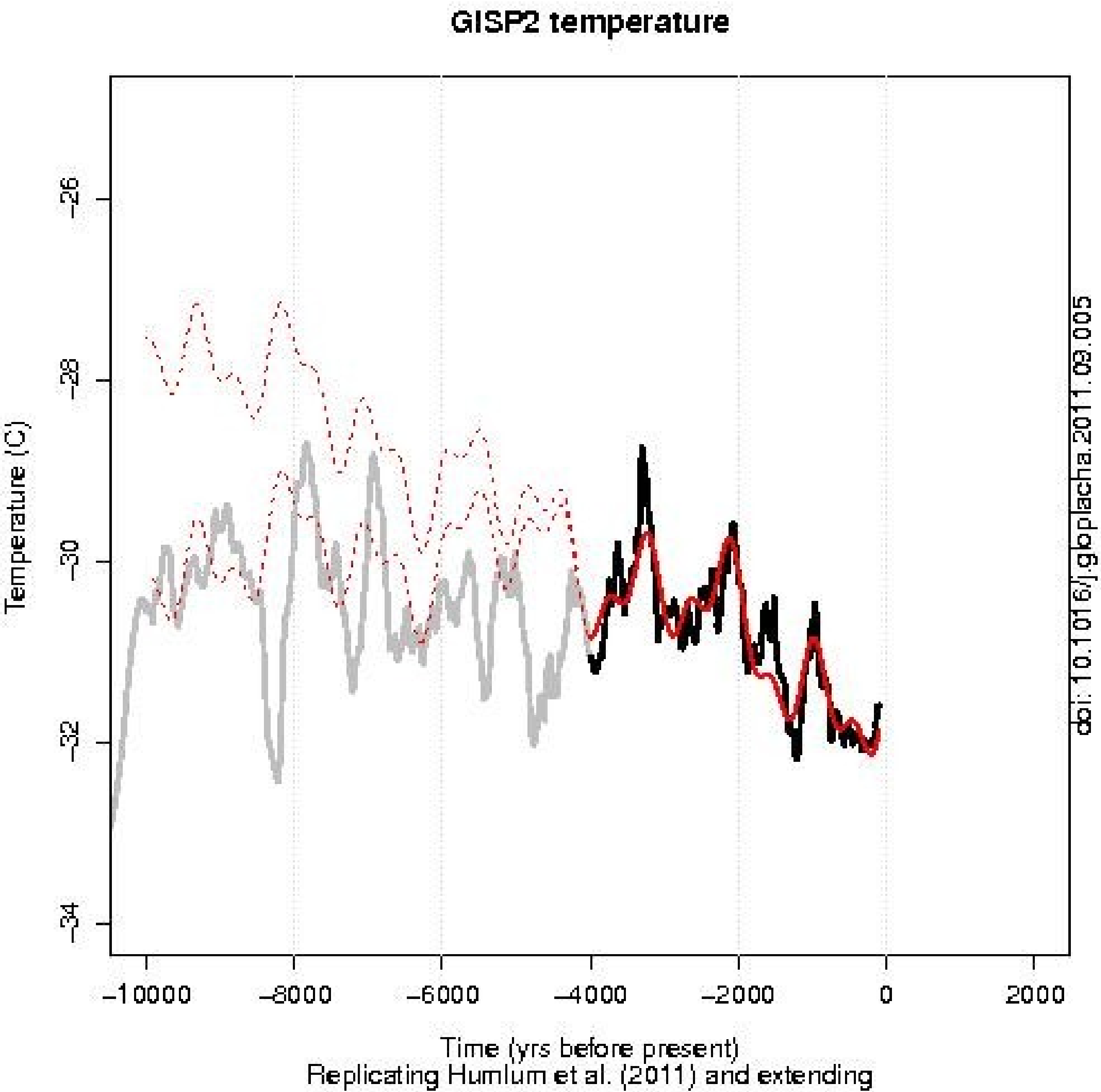
“We infer that the about 1130 and 590–560 year periods identified by us in the GISP2 core (Fig. 7) may correspond to the about 1000 and 500 year periods … “

“demonstrate how such persistent natural variations can be used for hindcasting and forecasting climate”

“Apparently the Moon may exercise a regional and global climatic control”.

“The climate is driven by Jupiter, Saturn and the moon”

Replication of prediction



Same data & frequencies

Some had been ignored

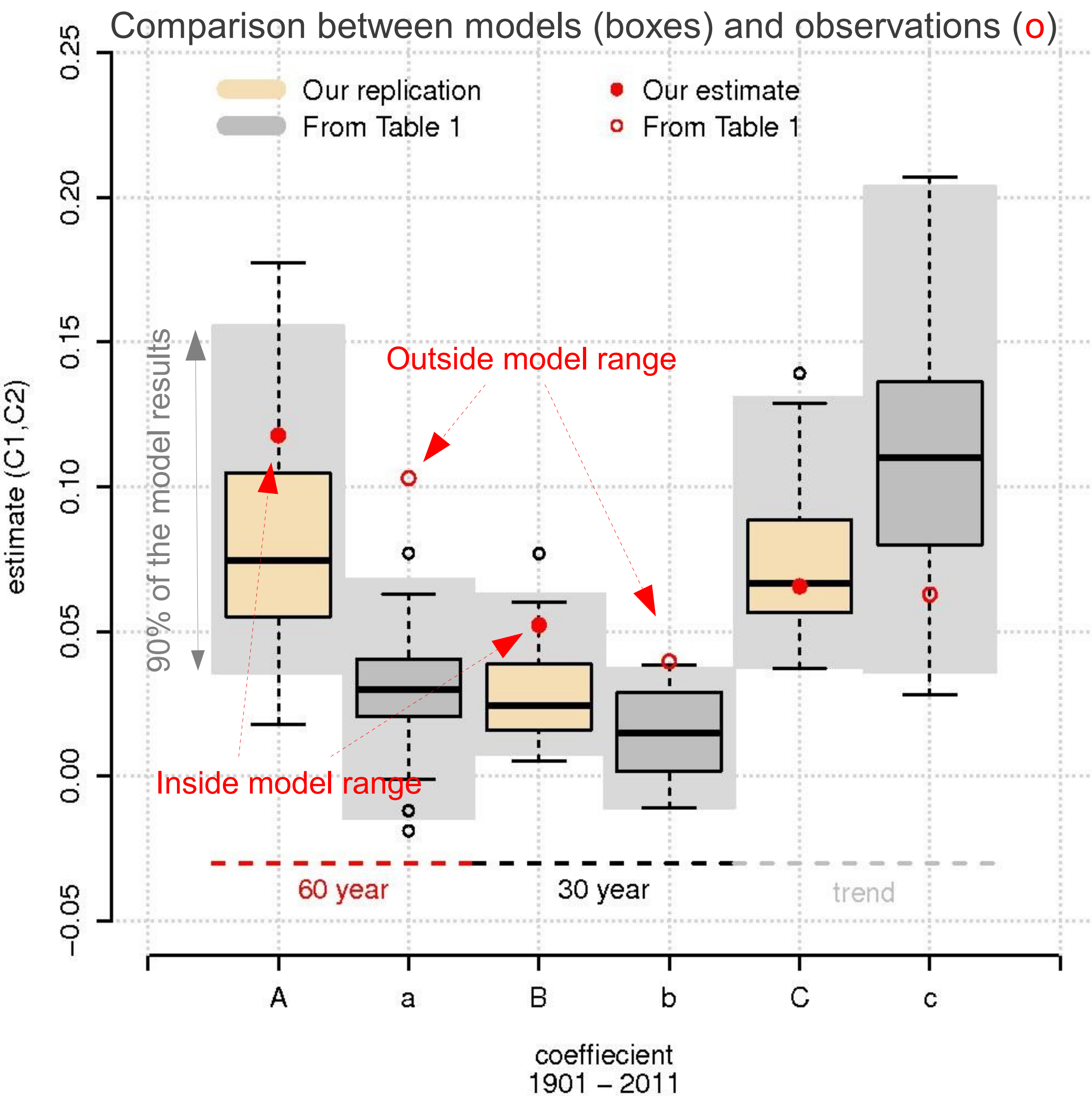
Extension of prediction

Model falsified

“The climate is driven by Jupiter, Saturn and the moon”

Case 3: Replication of previous tables

Scafetta (2011): Table 1



Phase in climate model results assumed constrained by great planets. Planets not accounted for in the models.

Same data & frequencies

Numbers copied from tables

Repeated analysis with correct statistics

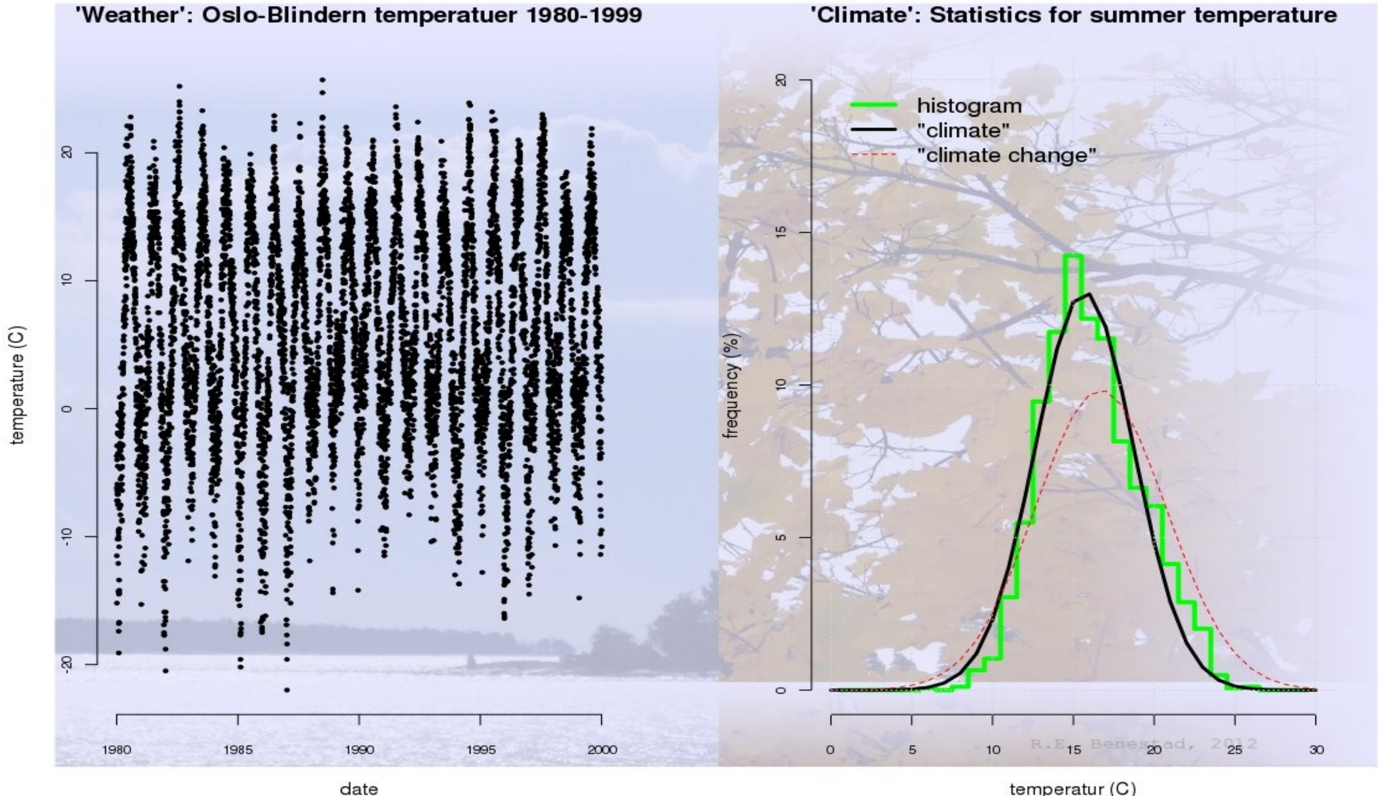
Objective model set-up

Data & models.

Verification of models & data.

The data

- Meta-data: sources!
- ReplicationDemos: 'attr(x,'URL')
- DOI & references.



Data

Measurements, observations.

Quality and quantity.

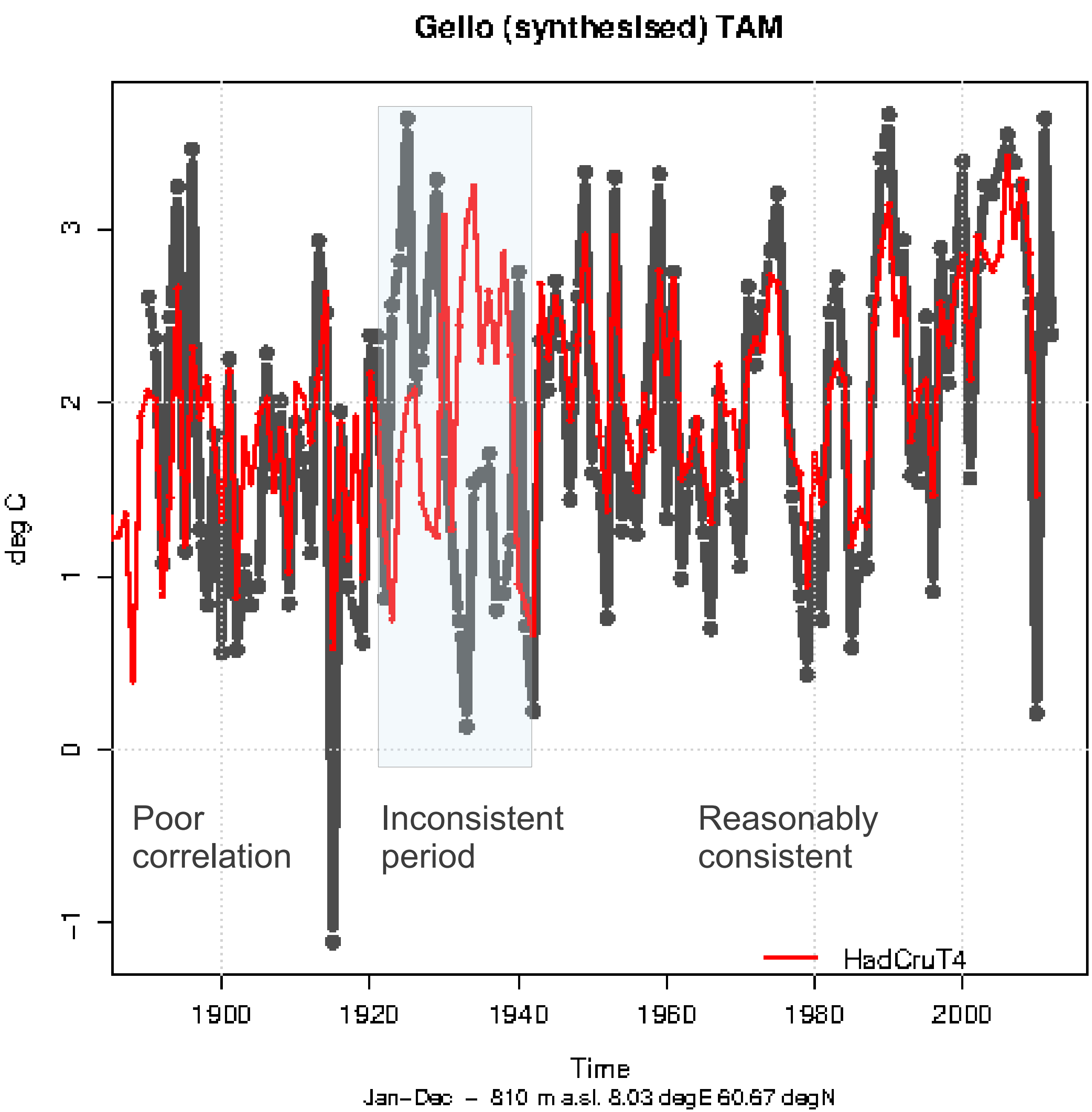
Meta-data: how were they measured and what do they really represent?

Errors & accuracy.

Hard to verify directly – measured on time...

Compare with other data and known situations.

Consistency – sample test

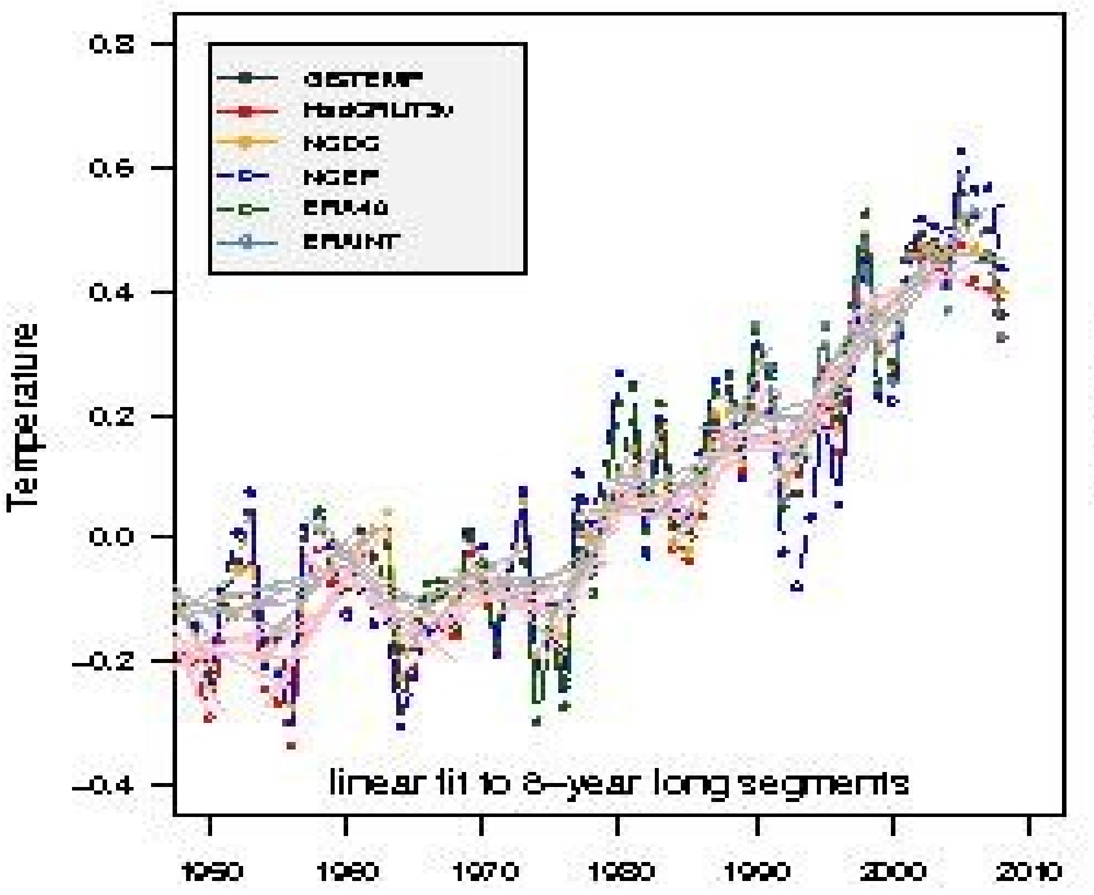


Interpolate annual mean temperature data from HadCRUT4 same coordinates as **Geilo.**

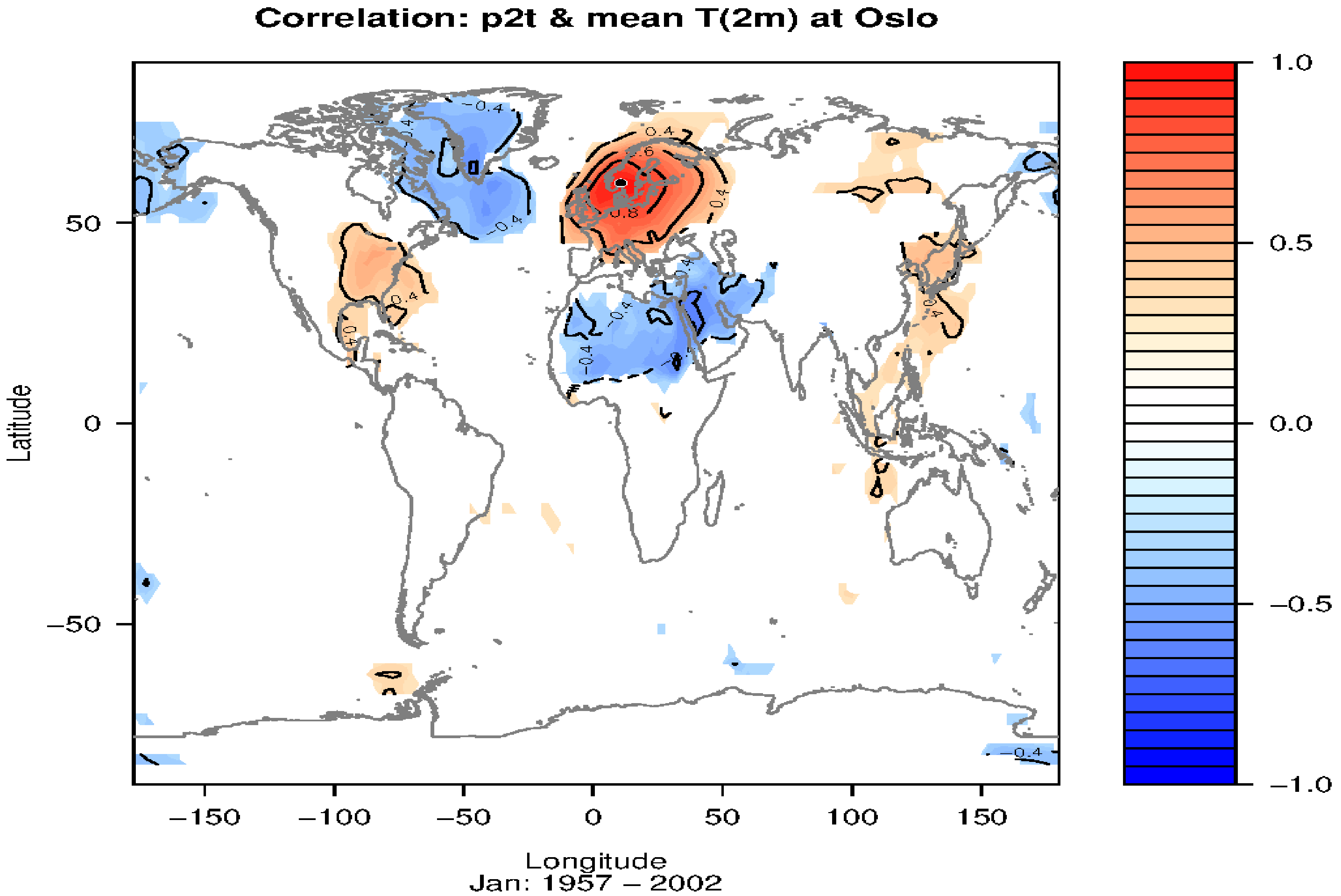
Geilo: many short sequences stiched together.

Different global mean analyses

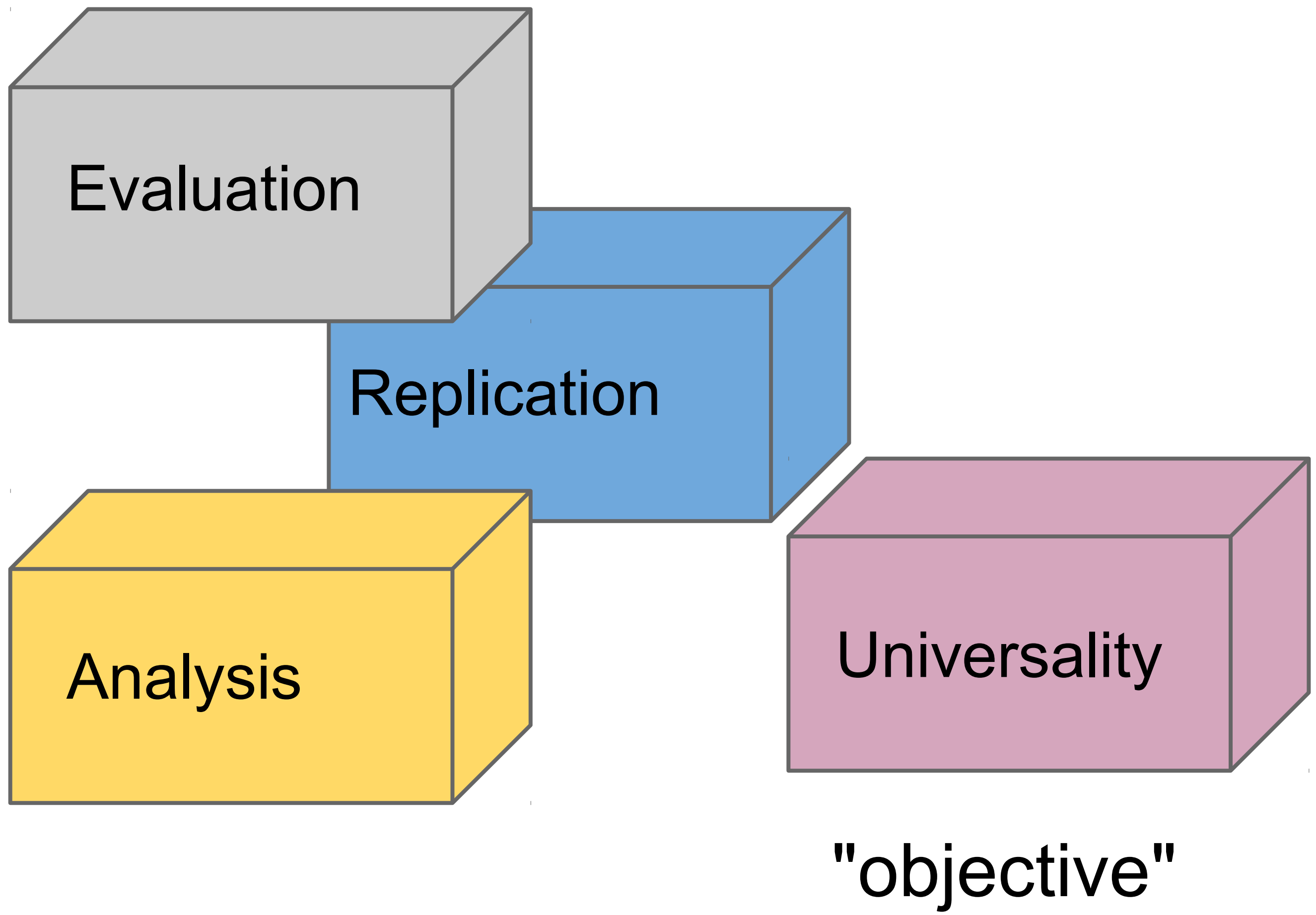
- Consistency between different analyses on trends.
- Observation & reanalyses.



Consistency



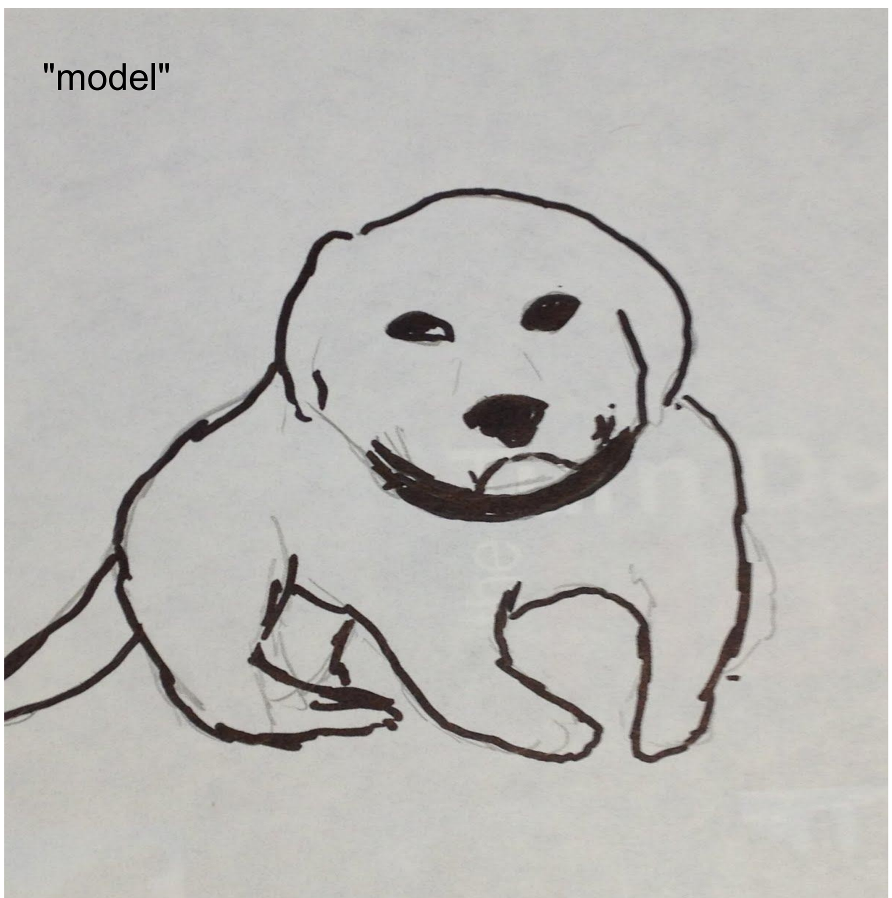
How good is my model?



What do we mean by a 'model'?

Purpose

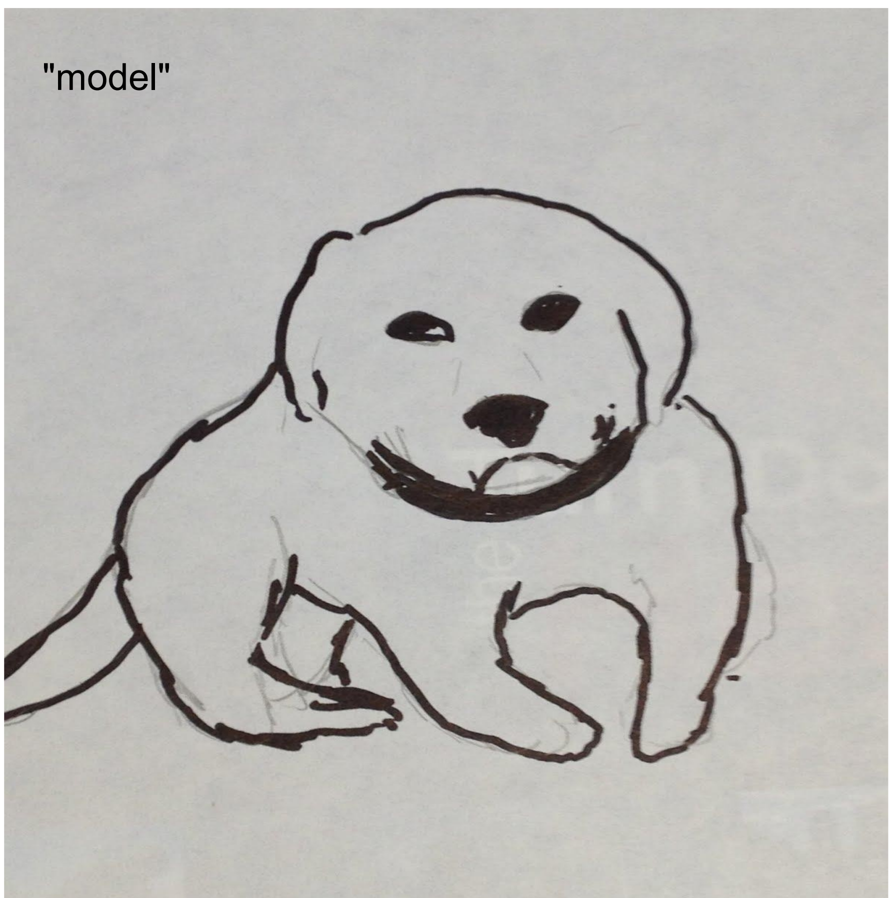
What information does it convey?

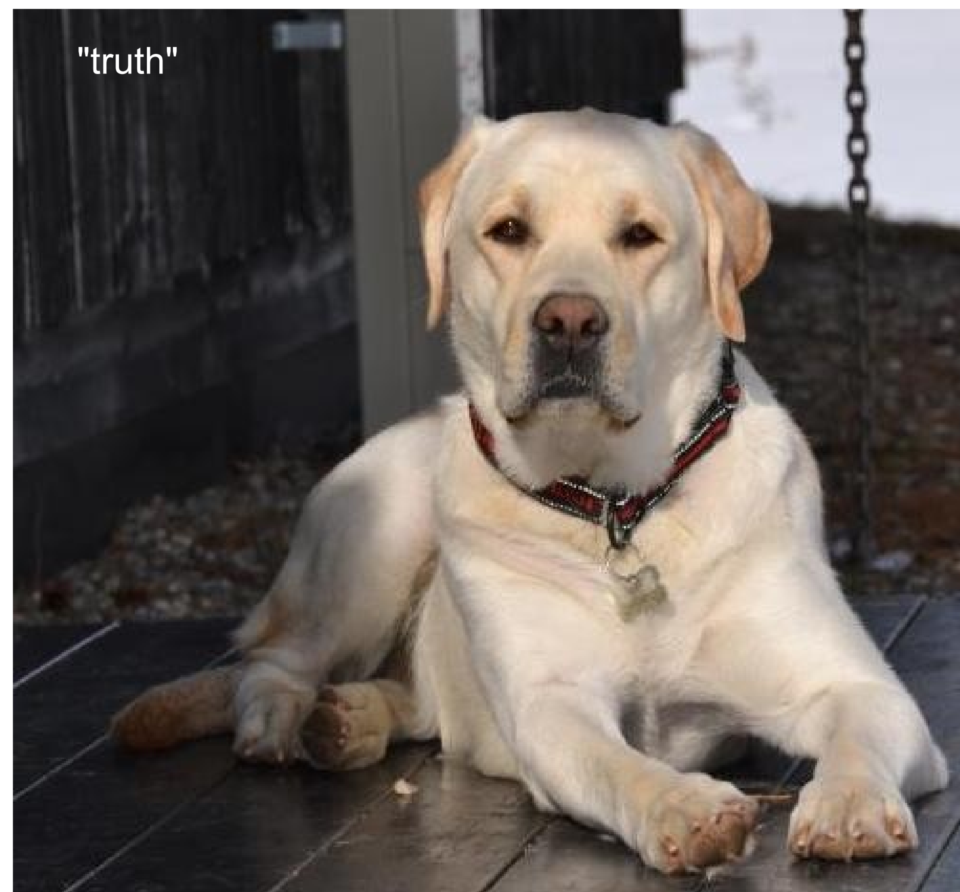


What do we mean by a 'model'?

Purpose

What information does it convey?

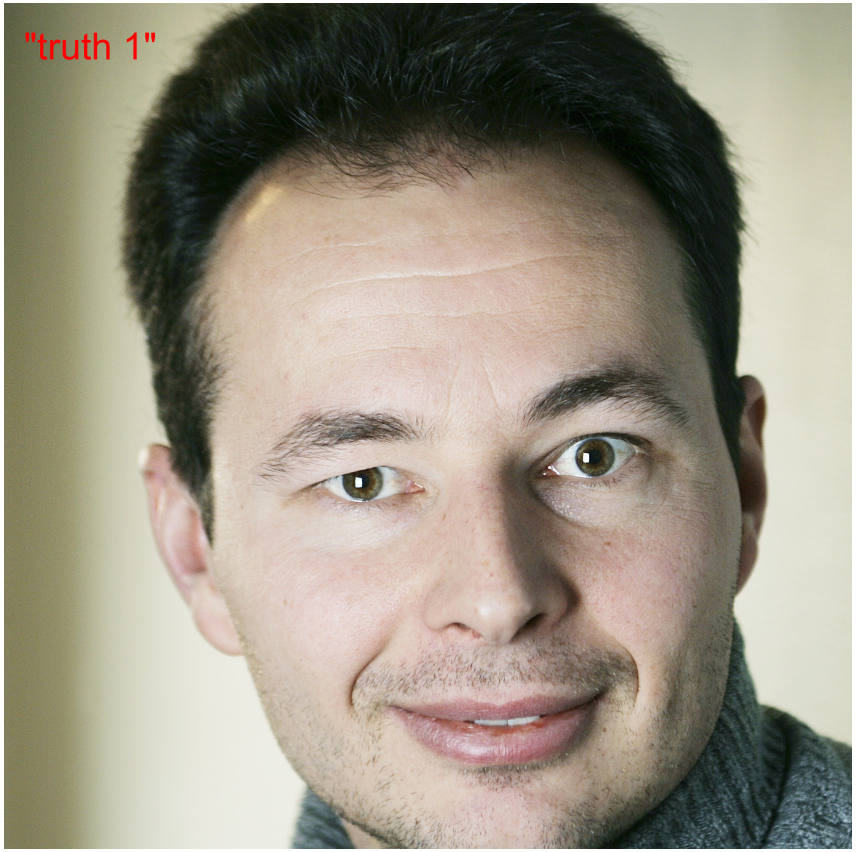




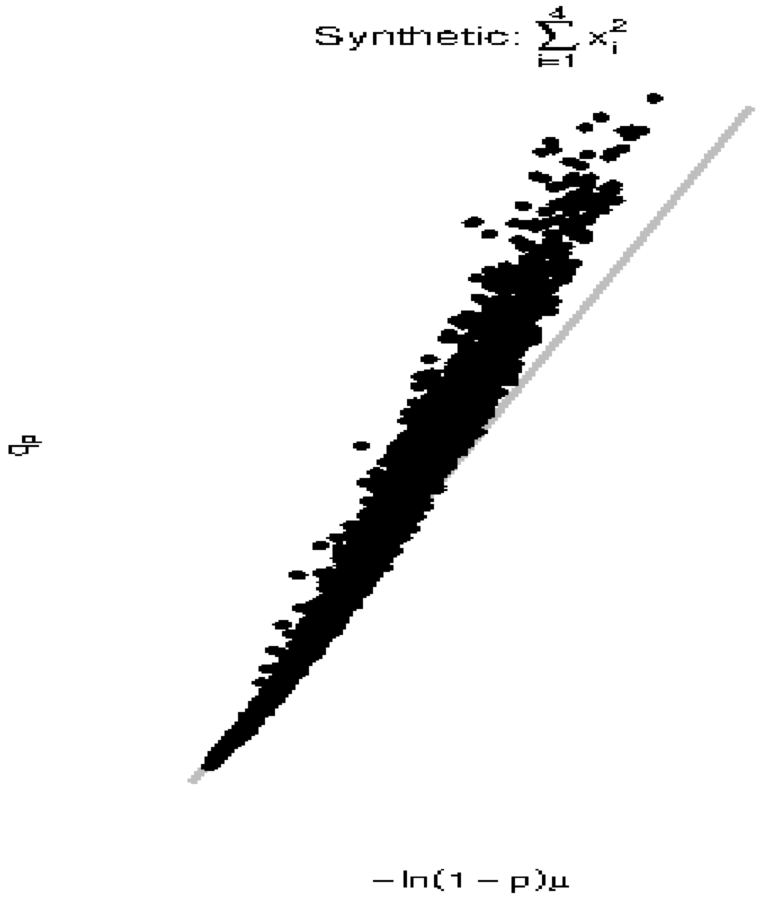
Which truth is closest??

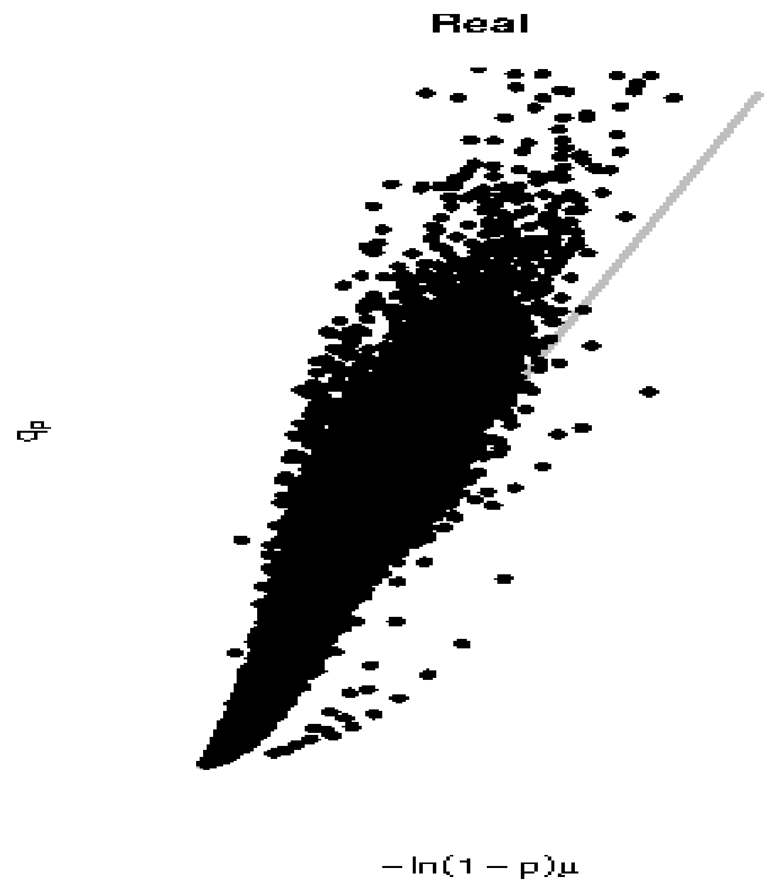
Purpose

hat information does it convey?

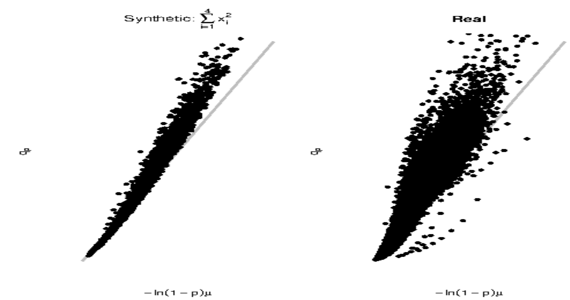


Similar features?



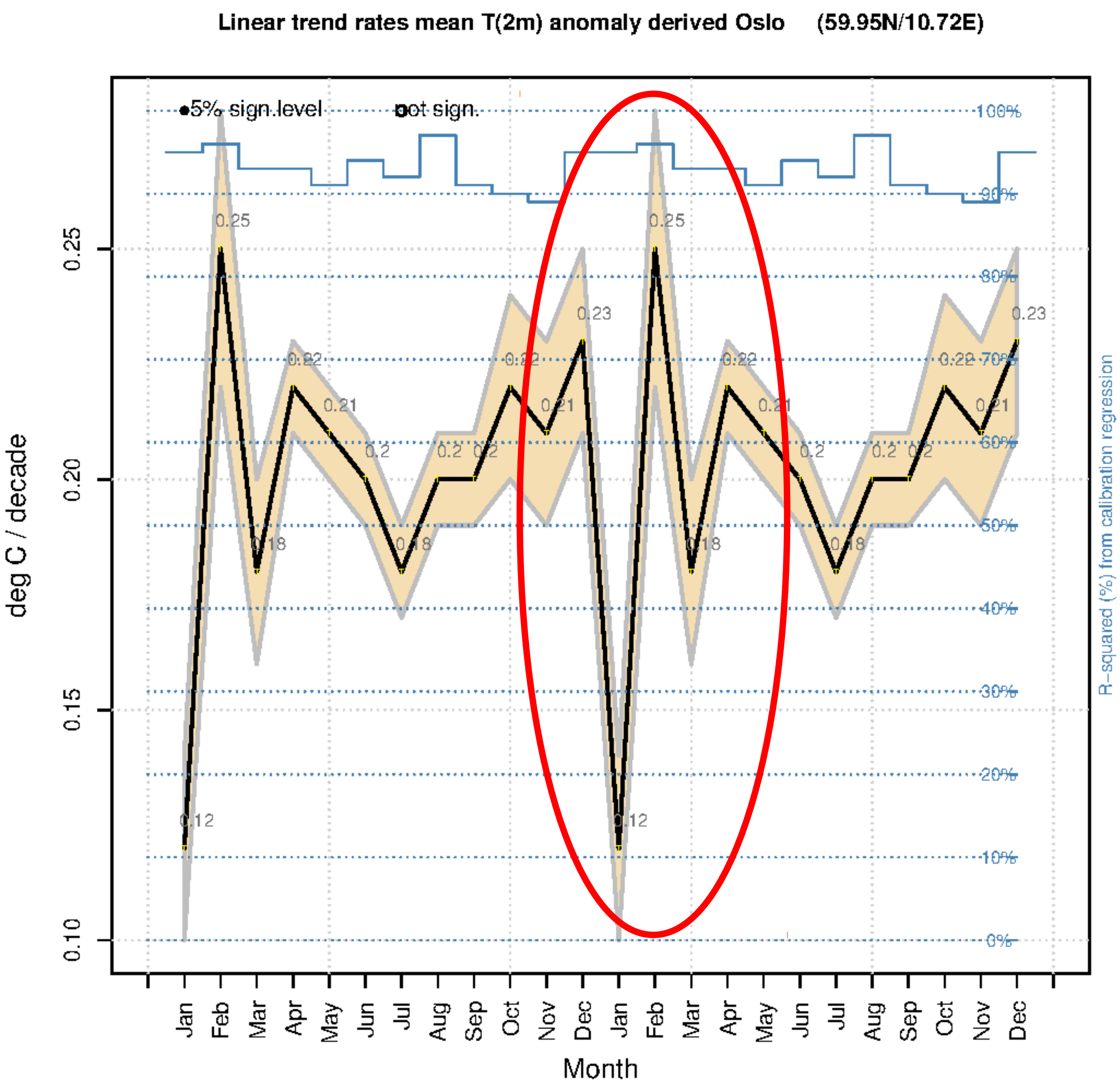


Not just the predictions



- More than just a set of numbers
- Diagnostics
- A range of diagnostics – look for consistency and realism – similarities...
  - Skill scores – treated more in detail later

Quality check – strange features?

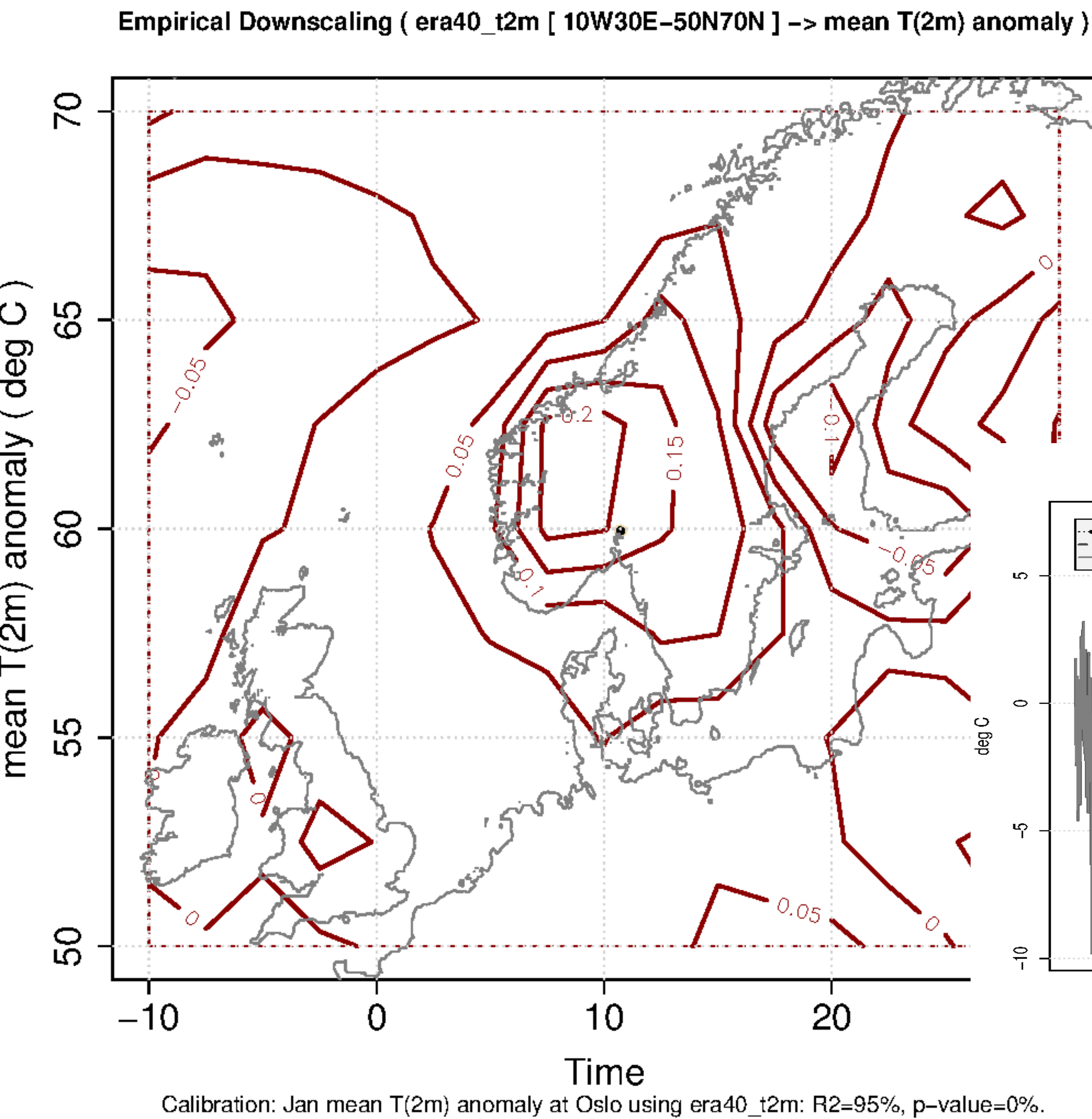


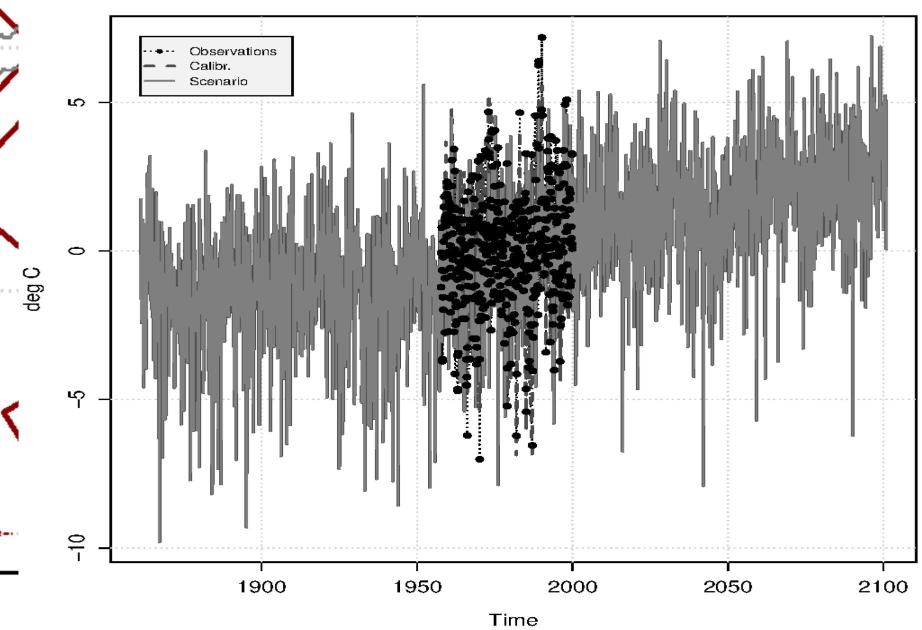
How predited trends vary through the season.

No reason to expect sharp and irregular jumps.

Smooth, simple, and slow functions ('Occam's razor').

Additional diagnostics





Dependence & independence

- 'Articifial skill' – picks information from the answer.
- Seperate data for calibration and data for testing.
- True model
  - Universially valid
  - Tough tests – extreme differences.
  - Objective

Avoid V&V on cherry picks

Double blinds avoid unconscious bias taint.

1st blind: e.g. subject taking the medicine

2nd blind: e.g. experimentalists is unaware of type of sample (medicine or placebo?).

Experimenter bias.

Harvard Univ. 1963 rat trials “bright” and “dull” from same stock. Borderline cases & selective abour recording.

Double blinds to avoid bias

1st blind: old observations not done for the specific purpose at hand

2nd: “blind injection” - add similar random samples (Monte-Carlo simulations)

Analyst unaware of which sample is which.

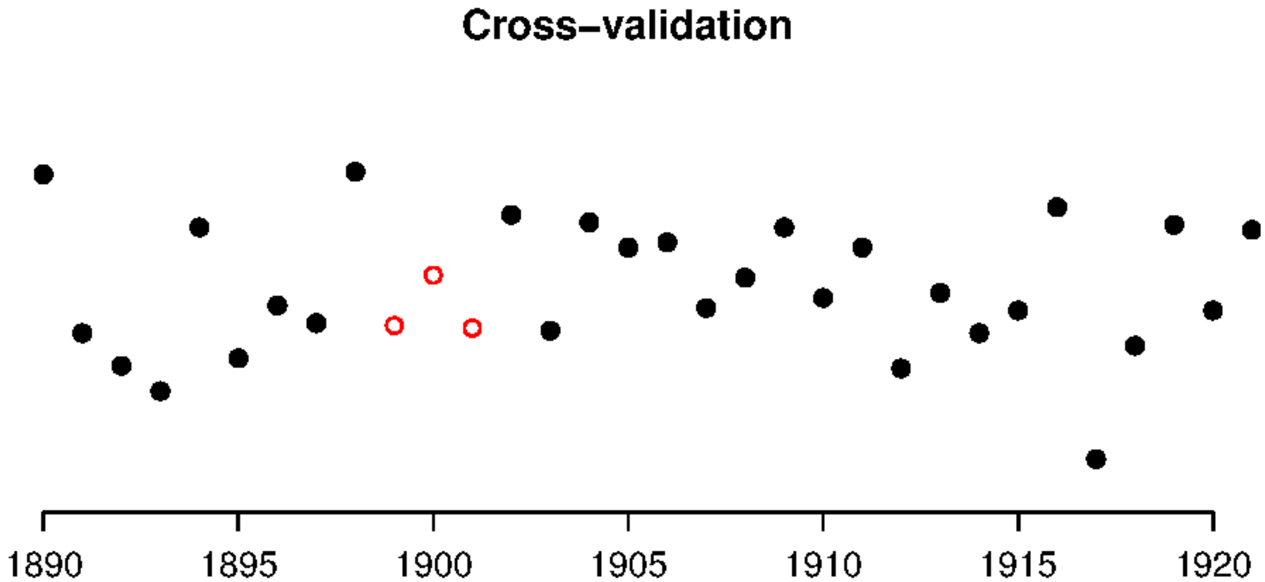
Brian Clegg (2013), 'The blind physicist',Physics World, January

Calibration: Cross-validation

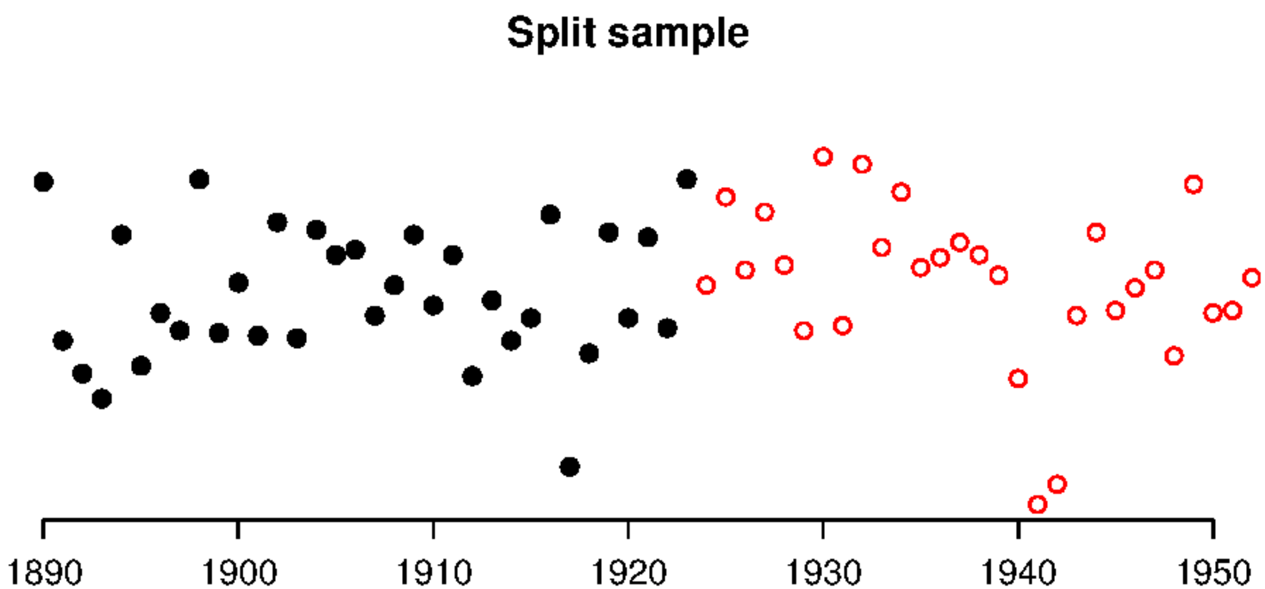
Potential problem: over-fit and fortuitous weighting giving accidental good match.

Solution: Split sample. long series.

Alternatively: Stepwise screening (stepwise regression), or a combination.



Short series
Auto-correlation?



Long series
Long-term trends

Input-based verification or by parts.

Design for testing – code in tests.
If the problem can be solved analytically for certain cases (inputs), then write functions to test these and compare with known analytical solutions.
Test different part if there are clear aspects that can be extracted.
Conservation of mass, energy, charge, etc. can be useful.

Next lecture