The Reliability of Millennial Multi-proxy Temperature Reconstructions

Workshop on Mathematics in the Geosciences
Northwestern University
October 3, 2011

Speaker: Blake McShane, Northwestern University
Coauthor: Abraham Wyner, University of Pennsylvania

What is Paleoclimatology?

- Measurements of climate variables such as temperature and precipitation typically exist only for the last one or two hundred years (the so-called Instrumental Period).
- Natural phenomena such as tree rings, ice cores, lake sediments, speleothems, and corals can be measured several hundred or even thousand years back in time.
- Such natural phenomena are called climate "proxies" because it is believed that they can serve as a proxy or surrogate for climate variables.
- Paleoclimatology is the *statistical* task of:
  - Building a model to predict climate variables from proxies using data from the Instrumental Period where both are observed.
  - Using the model in combination with the proxies to reconstruct the value of climate variables at times prior to the Instrumental Period.
- A particular area of focus is on *Northern Hemisphere Annual Average Temperature*.

The Data (Mann *et al.*, 2008)

- Today's Talk:
  - Y = {y_t}: the vector of Northern Hemisphere annual average land temperature for t=1850,...,1998 (i.e., years AD).
    Note: y_t is computed by the University of East Anglia Climatic Research Unit using an *unknown* area-weighted spatial average of the {z_{tj}} (see below).
  - X = {x_{tj}}: a matrix of natural proxies for t=998,...,1998; j=1,...,93.
- Additional Data:
  - X is actually a ragged array of 1209 natural proxies, each of which has its own starting date. For today, we focus only on the 93 which go back 1,000 or more years.
  - Z = {z_{tj}}: a matrix of local temperature given on a 5° longitude by 5° latitude grid for t=1850,...,1998; j=1,...,1732.
    Note: a 5° by 5° grid implies 2,592 cells in the global grid. Mann *et al.* disqualified 860 such cells because they contained less than 10% of the annual data thus leaving 1732.

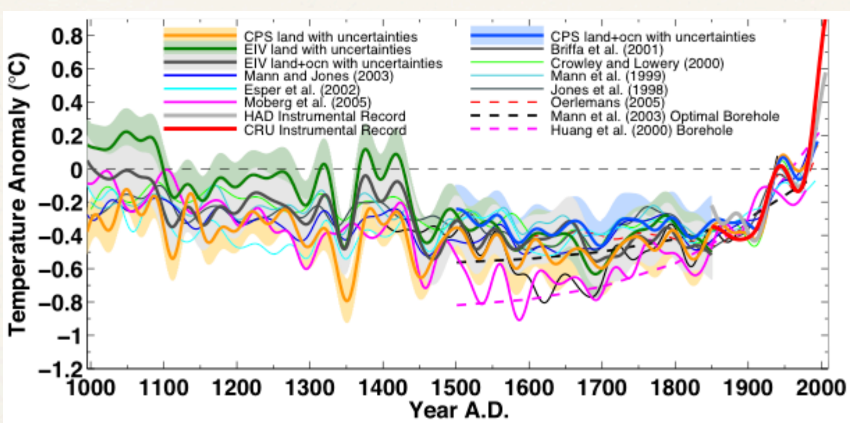
The Statistical Problem

- As noted, there are two parts:
  - Use the data for t=1850,...,1998 to model Y as a function of X.
  - Reconstruct or "backcast" y_t for t=998,...,1849 using the model and the X data for t=998,...,1849.
- There are also several major difficulties:
  - The data contain complex spatial and temporal dependencies (thereby making the effective sample size quite small).
  - The number of proxies is much larger than (or on the same order as) the number of instrumental years ($p >> n$ or $p \approx n$).

Are Such Reconstructions of NH Temperature Reliable?

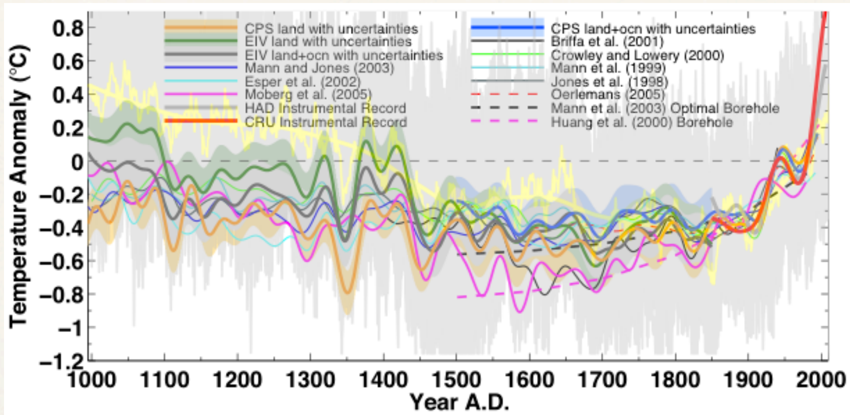
- Our work suggests ***NO*** based on three critical facts:
  - The proxies are unable to predict heldout blocks of instrumental temperature significantly better than random series generated independently of temperature.
  - Various models which perform similarly at predicting heldout blocks of instrumental temperature produce extremely different historical backcasts.
  - Our own model provides standard errors for the backcast (i.e., uncertainty intervals) which are extremely wide.

Reconstructions of Climate Scientists (Mann *et al.*, 2008)



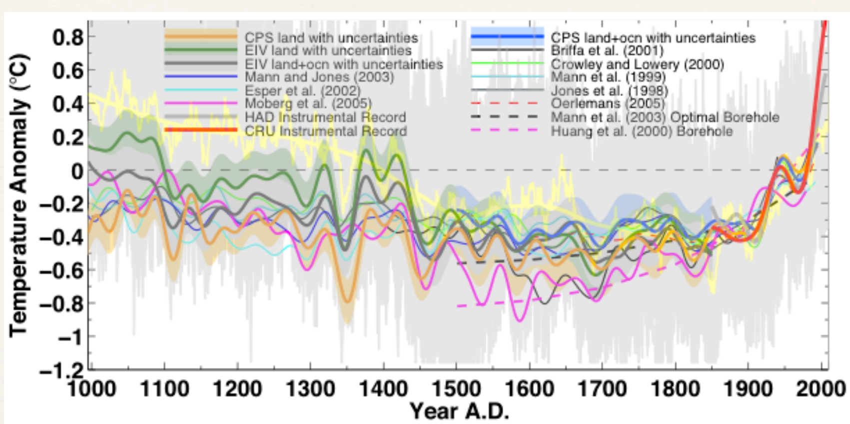
- Climate scientists have produced a variety of historical backcasts over the past decade.
- These backcasts tend to diverge as one goes farther back in time.
- Some of the backcasts also provide uncertainty intervals.
- RegEM EIV (Mann *et al.* 2008) given in green is considered to be among the best.

Our Model: Similar Backcast, Much Wider Intervals



- Our reconstruction is a Bayesian regression of NH Temperature on two lags of itself and the first ten principal components of the proxies (i.e., Bayesian AR2+PC10).
- It is given in yellow and largely matches the reconstructions of climate scientists (although it is somewhat warmer during the Medieval Warm Period).

Our Model: Similar Backcast, Much Wider Intervals



- But, our uncertainty intervals (gray) which are fully pathwise and account for parameter uncertainty are much wider than those published in the climate science literature.
- These intervals are wide enough to be consistent with all published reconstructions.
- Climate scientists' intervals are inconsistent with our model and each others' models.

Why the Disparity in Uncertainty Interval Width?

- Thirteen teams of statisticians, climate scientists, economists, and popular bloggers have prepared discussions of our work (*Annals of Applied Statistics*, March 2011).

- This discussion uncovered a number of interesting points which suggest reasons for the disparity in interval widths.

- We focus on two today:
  - "Your" methods are "bad" whereas our method (RegEM EIV) has proven itself powerful, particularly in GCM simulations (Schmidt, Mann, Rutherford).
  - GCM simulations must be assessed if they are to serve as a surrogate for controlled experiments of reconstruction methods (Berliner; McShane and Wyner).

What Are General Circulation Models (GCMs)?

- GCMs are large-scale mathematical models of the earth's climate used to simulate histories of temperature and other climate variables.

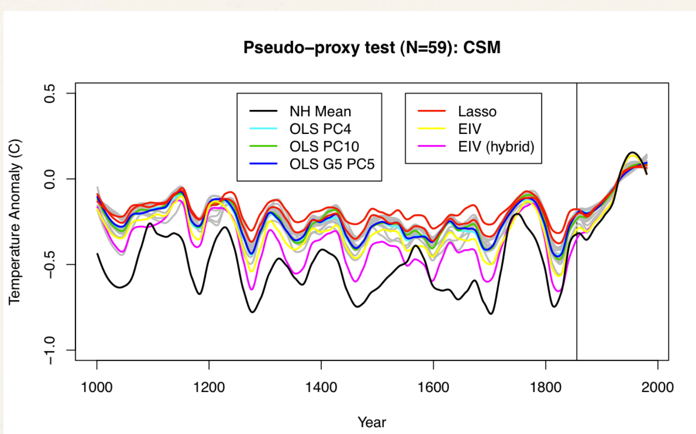
- The Output:
  - $Z = \{z_{tj}\}$: a matrix of local temperature given on a longitude/latitude grid.
  - $Y = \{y_t\}$: the vector of Northern Hemisphere average temperatures formed as an area-weighted spatial average of the $\{z_{tj}\}$.
  - $X = \{x_{tj}\}$: a matrix of "pseudoproxies" (i.e., $x_{tj} = z_{tj}$ + AR1 Noise).
  - One can set t and j at various levels of resolution.

- Again and as is typical, we will focus on Y and X at the annual level.

GCMs in Paleoclimatology

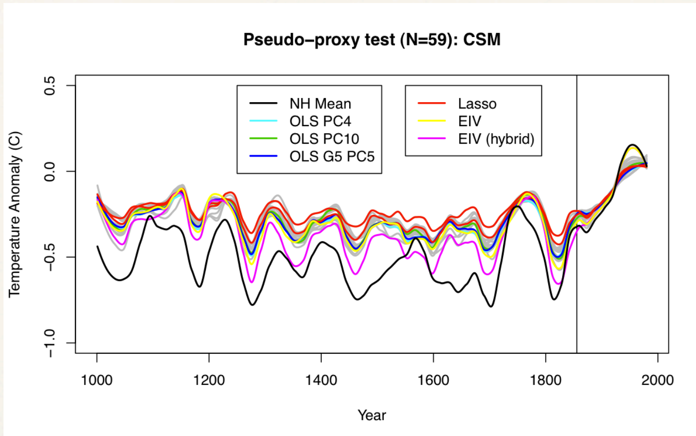
- GCM simulations have long been used to validate paleoclimatological models:
  - Schmidt, Mann, and Rutherford: “A standard benchmark in the field is the use of synthetic proxy data known as ‘pseudoproxies’ derived from long-term climate model simulations where the true climate history is known, and the skill of the particular method can be evaluated [see, e.g., Mann *et al*. (2007); Jones *et al*. (2009) and numerous references therein].”
- Principal simulations used include:
  - CSM: The National Center for Atmospheric Research (NCAR) Climate System Model.
  - GKSS: Helmholtz-Zentrum Geesthacht Research Centre European Centre Hamburg Ocean Primitive Equation-G (ECHO-G) “Erik” Model.
- The Validation Task:
  - Use a subset of the data (e.g., t=1856,...,1980) to model Y as a function of X.
  - Reconstruct or “backcast” y_t for the remaining time periods (e.g., t=1000,...,1855) using the model and the X data for t=1000,...,1855.
  - Compare the resulting reconstruction to y_t.

“Your” Methods Are “Bad” (Schmidt, Mann, and Rutherford)



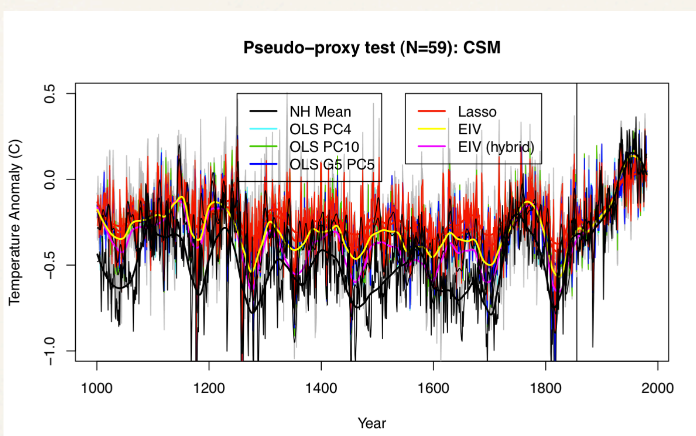
- RegEM EIV more closely matches NH temperature in the out of sample period than ordinary least squares regression fit to principal components of pseudoproxies or the Lasso fit to pseudoproxies.
- All reconstructions have roughly the same contours, though quite different levels.
- Does anyone notice anything funny about the *in sample* period?

"Our" Methods Are Not So "Bad"



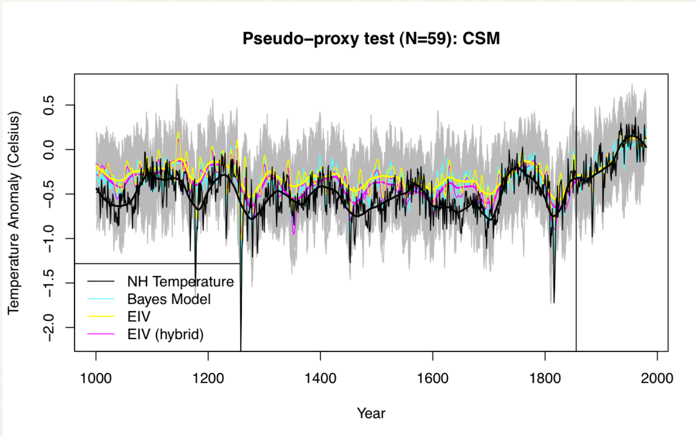
- OLS (and Lasso) had non-zero average residual in sample due to an incorrect, *ex post* "centering" step.
- When corrected, only Hybrid RegEM EIV substantially outperforms other methods.
- But, Hybrid RegEM EIV makes use of additional data thus "hybrids" can be constructed for all methods.
- Furthermore, has RegEM EIV been tuned to the simulation or vice versa?

Annual Variation of Model Fits Dwarfs Variation Across Models



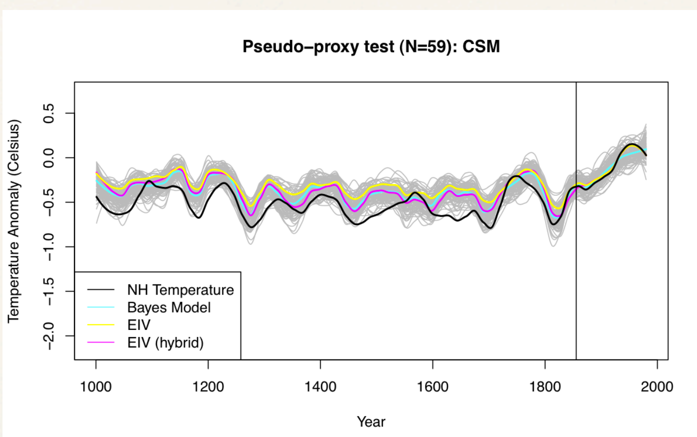
- When annual (rather than smoothed) backcasts are plotted, differences among the various methods appear inconsequential in comparison to the annual variation.
- This suggests that uncertainty intervals should be wide.
- Curiously, our actual method (i.e, Bayesian AR2+PC10) was omitted from these plots.

Our Bayesian AR2+PC10 Model:
Similar Reconstruction, Wide Intervals



- Wide intervals are necessary to accommodate annual variations in temperature.
- In sum, our method produces reconstructions that are practically equivalent to those of RegEM EIV (even to the hybrid version which uses additional data).
- But, it also provides properly calibrated uncertainty intervals.

Our Bayesian AR2+PC10 Model:
Still Calibrated When Smoothed



- Our wide intervals are necessary even when the data is smoothed.
- This, ***insofar as GCMs are useful for validation***, this suggests that our method is as accurate as the best methods in climate science and, more importantly, that our wide intervals for the real temperature and proxy data might indeed be correct.

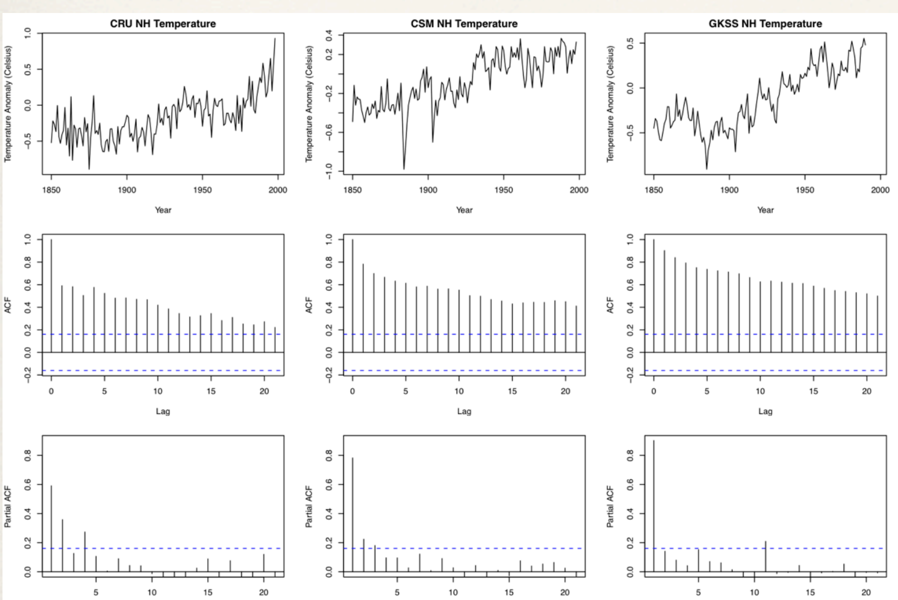
Are GCMs Useful for Validation of Paleoclimatological Models?

- Climate scientists, when evaluating these simulations, have focused on several technical issues ***internal*** to the simulation procedure:
  - Smerdon *et al*. (2008) shows that Mann *et al*. (2007) employed an inappropriate interpolation of GKSS temperatures and that verification statistics "are weakened when an appropriate interpolation scheme is adopted".
  - Smerdon *et al*. (2010) "identified problems with publicly available versions of model fields used in Mann *et al*. (2005) and Mann *et al*. (2007)" thereby showing that "the quantitative results of all pseudoproxy experiments based on these fields are either invalidated or require reinterpretation."

- Statisticians approach the evaluation of simulated data from a somewhat different and ***external*** perspective:

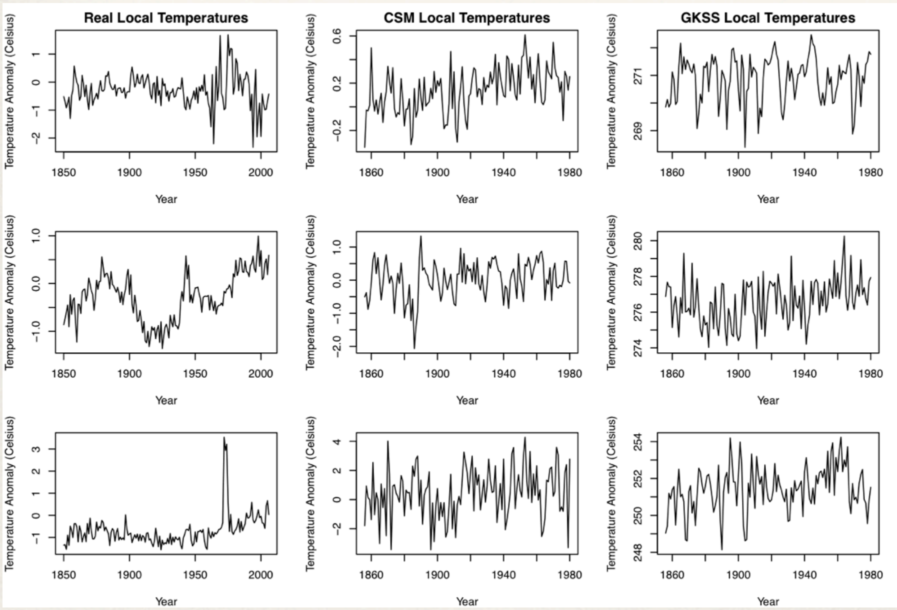
  If one wants insights gleaned from simulated data to carry over to real data, then key features of the simulated data should match key features of the real data.

Real NH Temperature Versus GCM NH Temperature



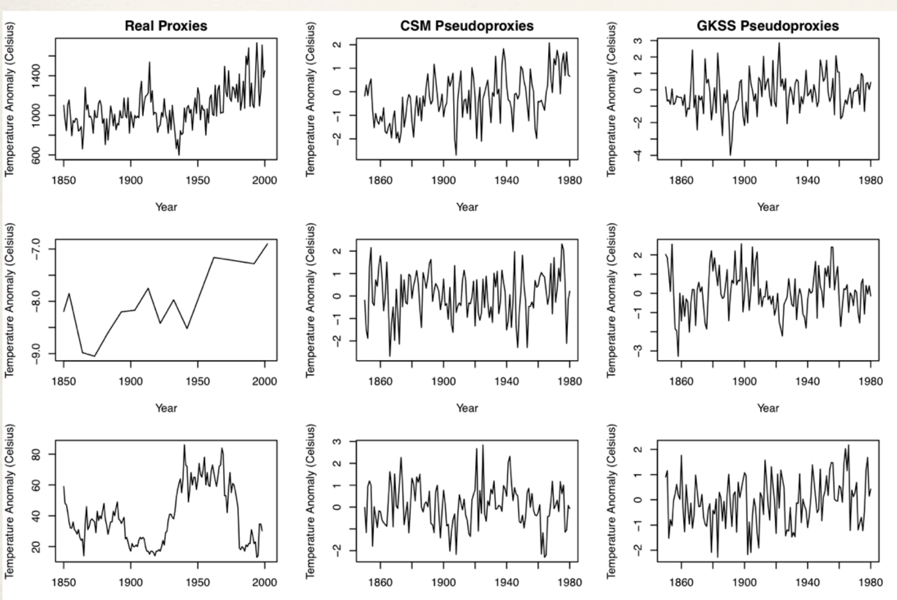
- Simulations have smoother NH temperatures and autocorrelation functions.
- Partial auto-correlations die out more quickly in simulations.
- Simulations seem to have only one or two distinct segments, unlike the "three or possibly four segments" in real temperature discerned by Davis and Liu.

Real Local Temperatures Versus GCM Local Temperatures



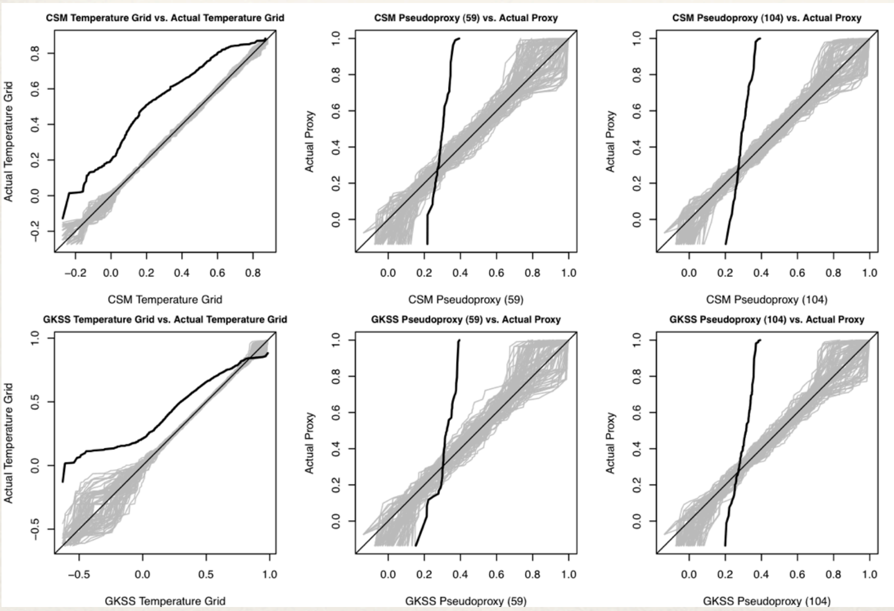
- Real local temperatures take many shapes:
  - Spitsbergen island in the Svalbard archipelago in the Arctic.
  - The north portion of the Omsk oblast in southwestern Siberia.
  - Baysuat in the Aktobe Province, Kazakhstan.
- GCM local temperatures do not vary as much across time and space.

Real Proxies Versus GCM Pseudoproxies



- Real proxies also take many shapes:
  - Tree Rings in Montana.
  - Monsoons. in India.
  - Speleothems in Scotland.
- Pseudoproxies from GCMs vary less across time and across different series.

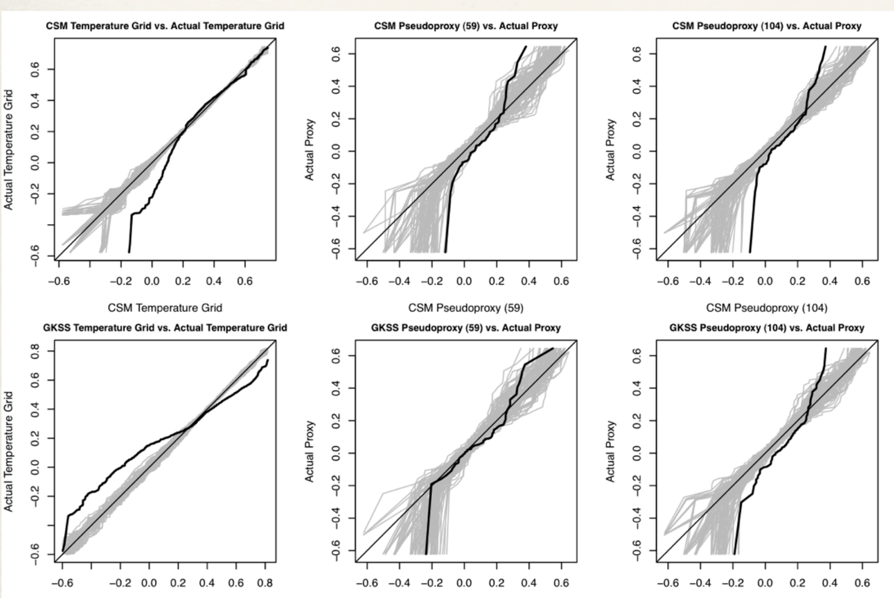
Lag One Correlation of Local Temperatures and (Pseudo)Proxies



- Bootstrapped QQ Plots show that actual local temperatures and proxies have lag one correlation coefficient distributions that are strikingly different than those from those of simulations.

- Smerdon and Kaplan (2007): "colored noise models used in pseudo-proxy experiments may not fully mimic the nonlinear, multivariate, nonstationary characteristics of noise in many proxy series."

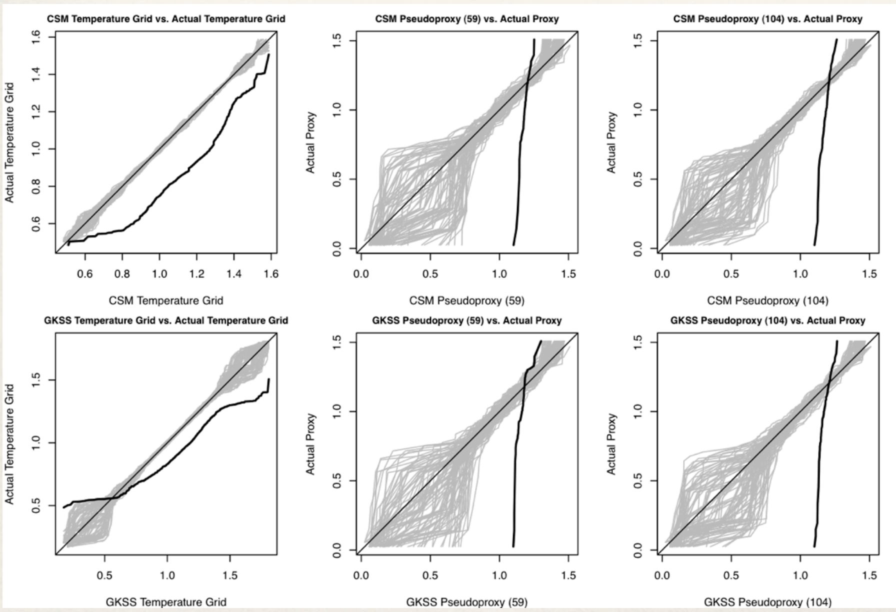
Correlation of Local Temp. and (Pseudo)Proxies with NH Temp.



- In general, the black QQ line teeters on or exceeds the thresholds set by the null bands.

- That is, the real data and GCM simulated data are unlike in terms of correlation with NH Temp.

“Smoothness” of Local Temperatures and (Pseudo)Proxies



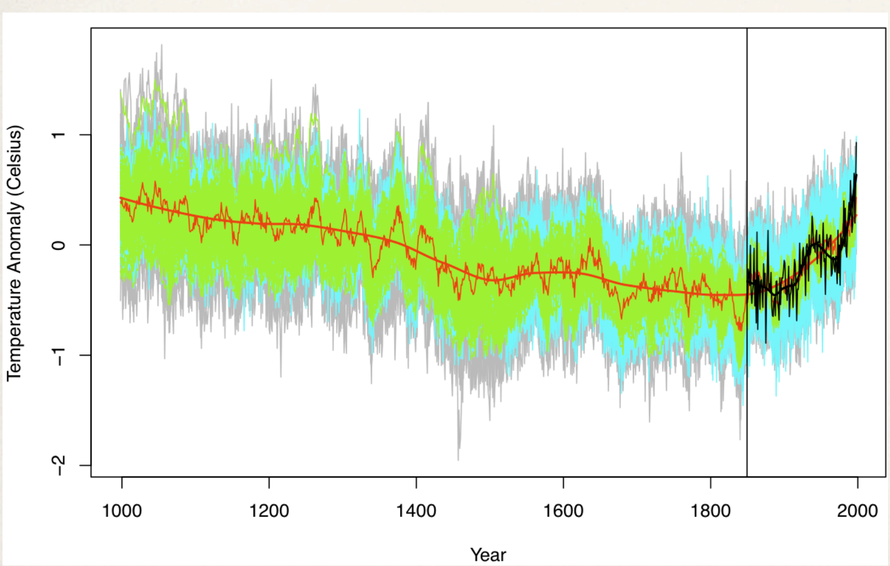
- The sample standard deviation of the first difference of each series (standardized before differencing).
- Again, the real data and the GCM simulated data fail to align.

Conclusions

- Our model provides reconstructions that are similar to those considered best by climate scientists--both on real data and on GCM simulated data.
- But our model produces extremely wide uncertainty intervals and their width is thoroughly validated by GCM simulations, a mainstay technique in the climate science literature.
- Nonetheless, important and obvious features of the real data are not replicated in the GCM simulated data.
- How applicable are GCM simulation results to the real data when such prominent features fail to match?
- Indeed, this should make us ***increase*** our level of uncertainty (indeed perhaps our wide intervals are in fact optimistically too narrow)!

Back Matter

Real Data: Variance Decomposition of Our Bayesian AR2+PC10 Model

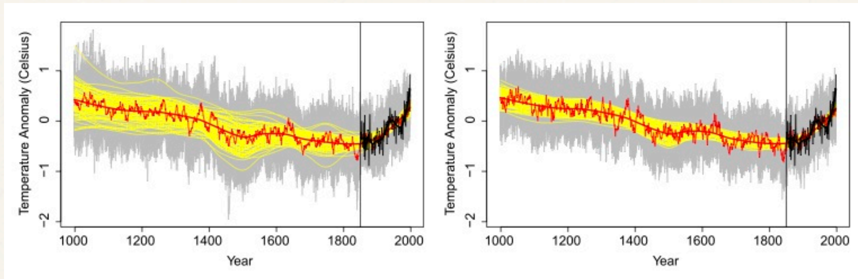


- **CRU NH Temp**.
- **Model Estimate**.
- **ε_t Uncertainty**.
- **Parameter Uncertainty**.
- **Total Uncertainty**.

- Not accounting for parameter uncertainty and the pathwise nature of reconstructions (both common in climate science) leads to overconfidence.

Real Data: Accounting For Temperature's Ability to "Self-Predict" Is Vital for Producing Properly Calibrated Confidence Intervals

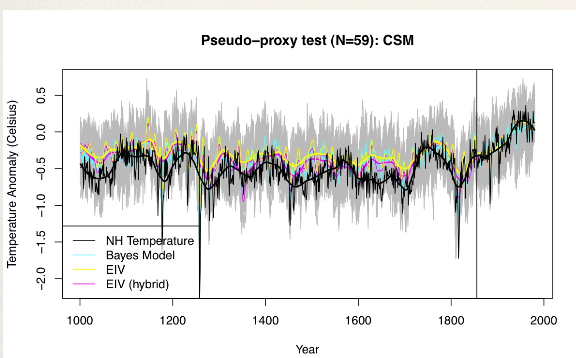
Raw and Smoothed Confidence Intervals

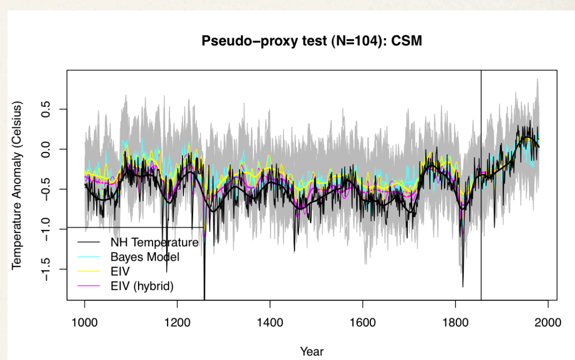


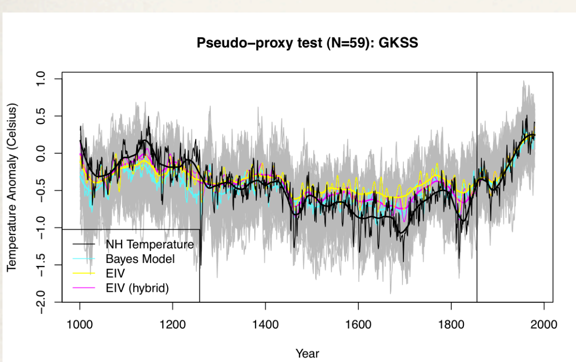
Accounting for Self-Prediction **Not Accounting for Self-Prediction**

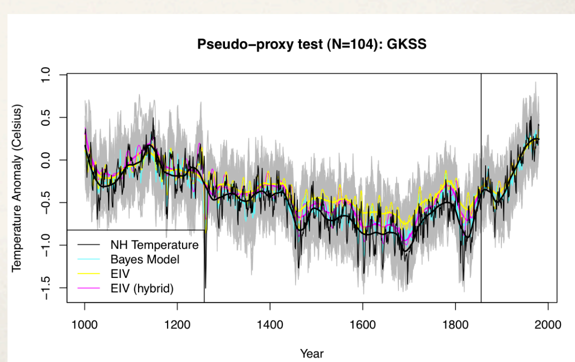
- In reconstructing temperatures and in validating reconstructive methodologies via pseudoproxy experiments, climate scientists typically do not account for temperature's ability to predict itself.
- Doing so is vital and has a large impact on confidence interval width--particularly when the intervals are smoothed as is common in climate science.

GCM Simulated Data: Calibration of Our Bayesian AR2+PC10 Model









GCM Simulated Data: Calibration of Our Bayesian AR2+PC10 Model