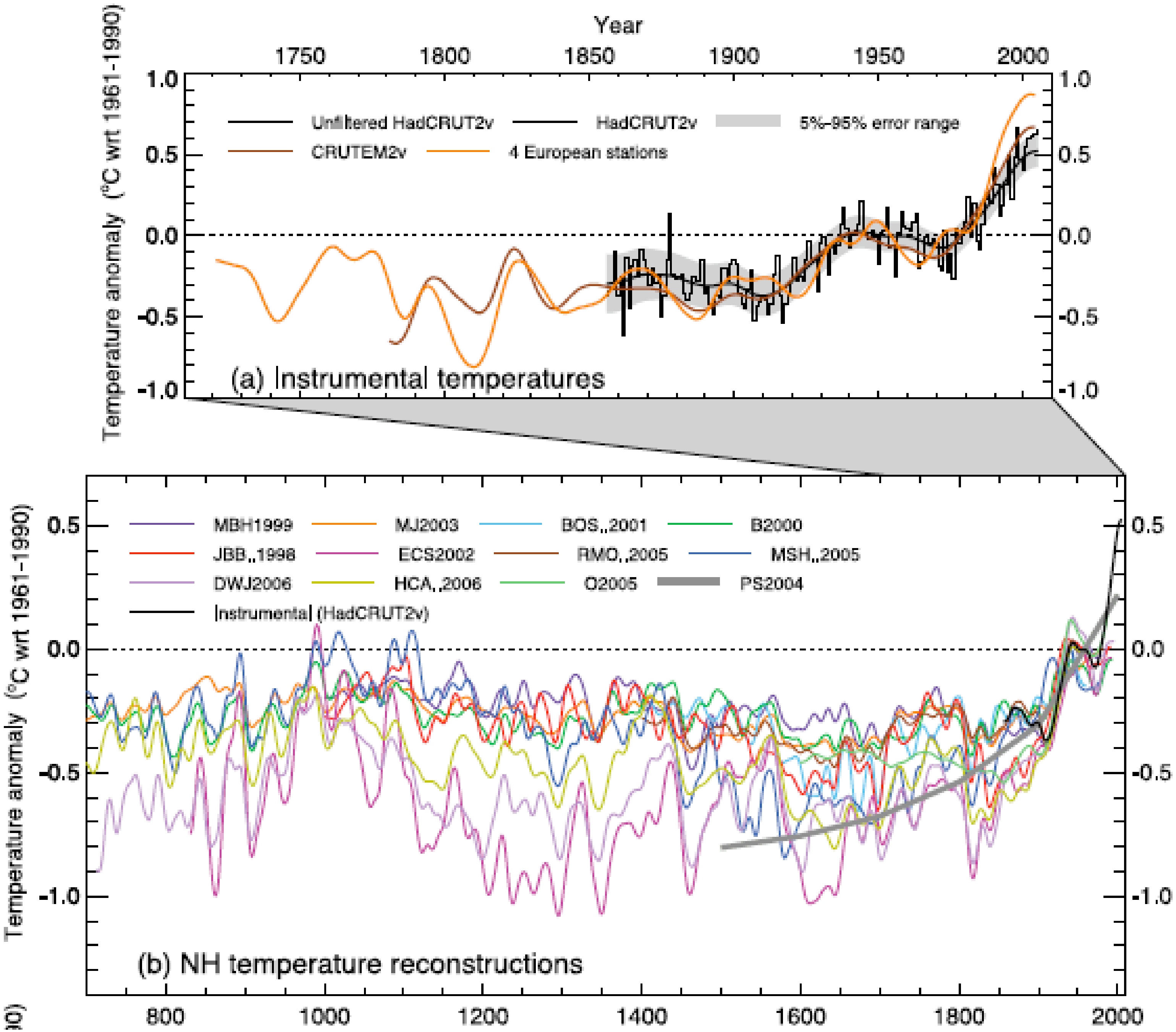
Estimating climate variables and covariances from incomplete data

Tapio Schneider
California Institute of Technology

Temperature estimates ("reconstructions")



Source: IPCC 2007

Incomplete data problems

- Estimation of temperature values from proxies is incomplete data problem:

  Given proxies x_a and relations between proxies and temperatures x_m estimable from period of overlap, what are estimated temperatures $\hat{x}_m$ in the past?

- Usually linear models (one per record) are used

$$x_m = \mu_m + (x_a - \mu_a)B + \varepsilon$$

with B estimated from period of overlap:

$$\hat{x}_m = \hat{\mu}_m + (x_a - \mu_a)\hat{B}$$

Some straightforward points

- (Co-)variances: sample variance of imputed values $\hat{x}_m$ underestimates variance of x_m (need to add variance of imputation error ε)
- Regression coefficient B depends on covariance matrix of x_m, which depends on missing values (nonlinear problem)
- Model based on assumption of missigness at random (may be violated in climate change context)

Expectation-maximization (EM) algorithm

1. Take estimated mean values $\hat{\mu}_{a,m}$ and covariance matrix $\hat{\Sigma}$ as given and compute $\hat{B} = \hat{\Sigma}_{aa}^{-1}\hat{\Sigma}_{am}$ and $\hat{x}_m = \hat{\mu}_m + (x_a - \mu_a)\hat{B}$ from them

2. Re-estimate $\hat{\mu}_{a,m}$ and $\hat{\Sigma}$ from completed dataset and from estimate of imputation error covariance $\hat{C} = \hat{\Sigma}_{mm} - \hat{\Sigma}_{ma}\hat{\Sigma}_{aa}^{-1}\hat{\Sigma}_{am}$

3. Iterate (1) and (2) until convergence.

EM algorithm converges monotonically but slowly (Dempster et al. 1977)

Regularized EM algorithm

1. Take estimated mean values $\hat{\mu}_{a,m}$ and covariance matrix $\hat{\Sigma}$ as given and compute regularized estimate $\hat{B}_r$ and $\hat{x}_m = \hat{\mu}_m + (x_a - \mu_a)\hat{B}_r$ from them

2. Re-estimate $\hat{\mu}_{a,m}$ and $\hat{\Sigma}$ from completed dataset and from estimate of imputation error covariance $\hat{C}$

3. Iterate (1) and (2) until convergence.

Regularized EM algorithm is only assured to converge (slowly) in special cases (e.g., Tikhonov $\hat{B}_r$ with fixed regularization parameter/prior variance).

Regularization I: Truncated total least squares

- Based on eigendecomposition of $\hat{\Sigma}$ with eigenvector matrix T :

$$\hat{B}_r = (T_{ar}^{+})^T (T_{mr}^T)$$

- Orthogonal regression of variables with missing values on variables with available values truncated at rank r.
- Takes errors in variables into account (symmetric in available and missing values): solves

$$\min \| [x_a - \mu_a, x_m - \mu_m] - [\hat{x}_a - \hat{\mu}_a, \hat{x}_m - \hat{\mu}_m] \|^2$$

$$\text{s.t.} \quad \hat{x}_a = x_a + \hat{\eta}, \quad \hat{x}_m - \hat{\mu}_m = (\hat{x}_a - \hat{\mu}_a) B$$

- Fast: only one eigendecomposition per iteration necessary

Fierro et al. 1997

Regularization II: Tikhonov regularization/ridge regression

- Regularization of $\hat{\Sigma}$ by addition of diagonal matrix

$$\hat{B}_r = (\Sigma_{aa} + h^2 D)^{-1}\Sigma_{am}, \quad D = \mathrm{diag}(\Sigma_{aa})$$

- In standard form, $\hat{B}_r = V\mathrm{diag}(f_j)\Lambda^+ F, \quad f_j = \lambda_j^2/(\lambda_j^2 + h^2)$, where $\Sigma_{aa} = V\Lambda^2 V^T$ (Wiener filtering of Fourier components F).

- Also takes errors in variables into account (arises as regularization of TTLS if relative error homogeneous; Golub et al. 2000)

- Slower: requires one eigendecomposition per record with missing values and per iteration

Schneider 2001

Regularization III: Choice of regularization parameter

- If principal interest is prediction (imputation of missing values), generalized cross-validation suggests itself
- Straightforward computationally with Tikhonov regularization (only scalar optimization necessary)
- But note: GCV function has mass point at zero (Wahba and Wang 1995), so regularization parameter must be bounded away from zero
- GCV also possible with TTLS, but more complicated computationally (van Huffel et al. 2006)

Schneider 2001

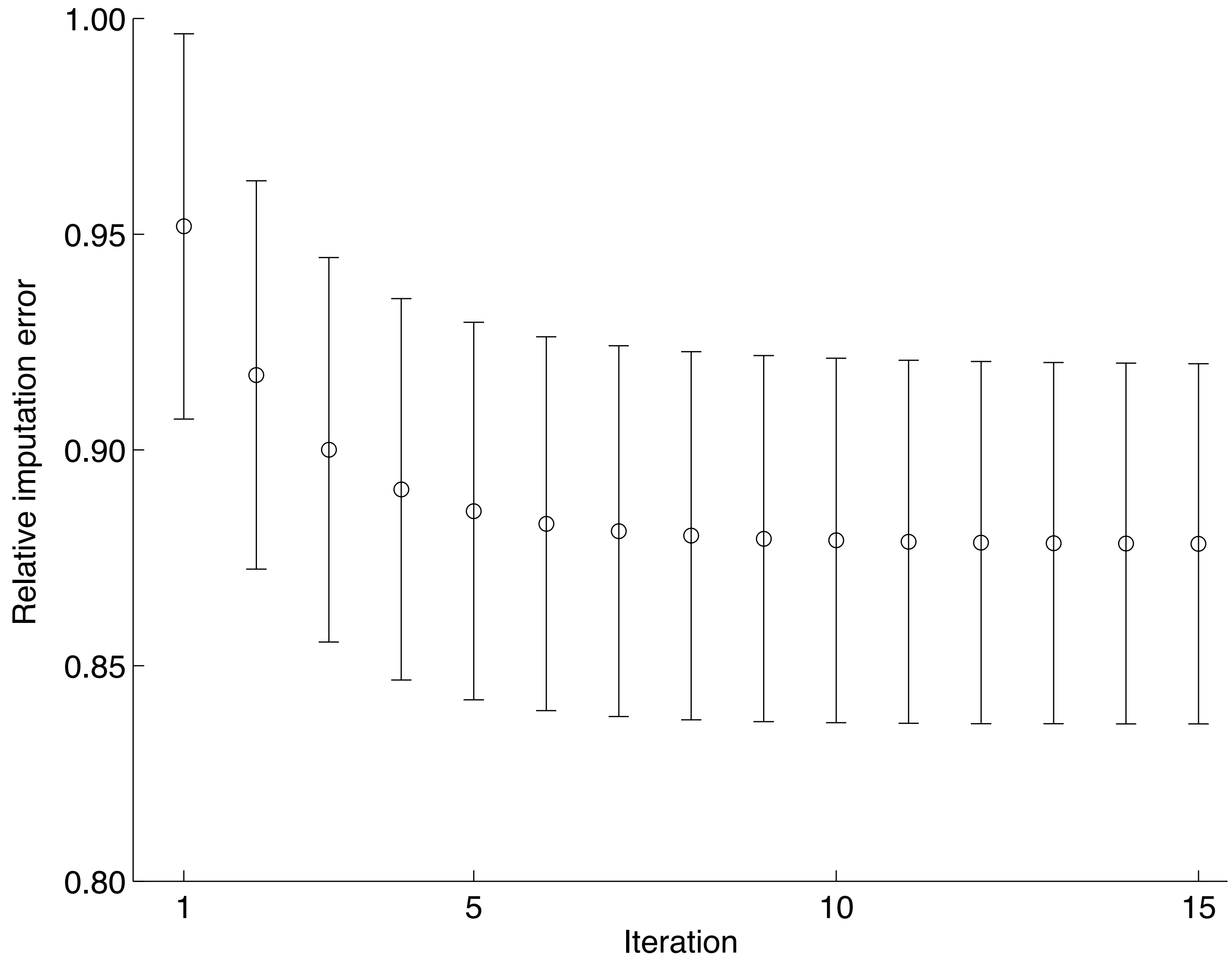
An example algorithm

- Use Tikhonov regularization/ridge regression with a separate regularization parameter estimated *for each* missing value
- Regularization parameter estimated by GCV (bounded away from zero using a discrepancy principle)
- Empirically, algorithm converges reliably
- Temporal covariance information can also be exploited

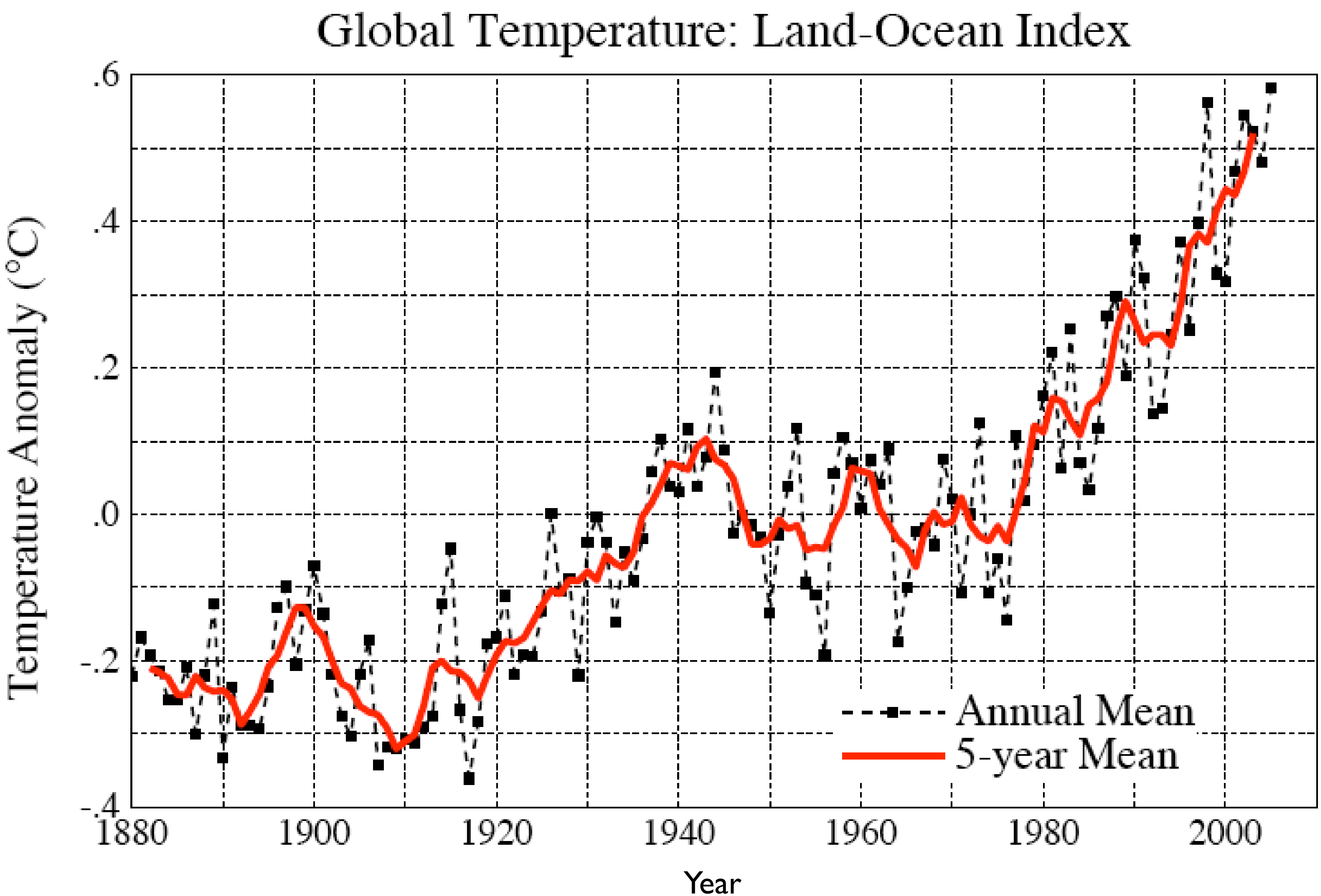
Systematic tests with realistic test data using different regularization approaches would be desirable

Schneider 2001

Convergence of regularized EM algorithm with ensemble of GCM simulated temperatures



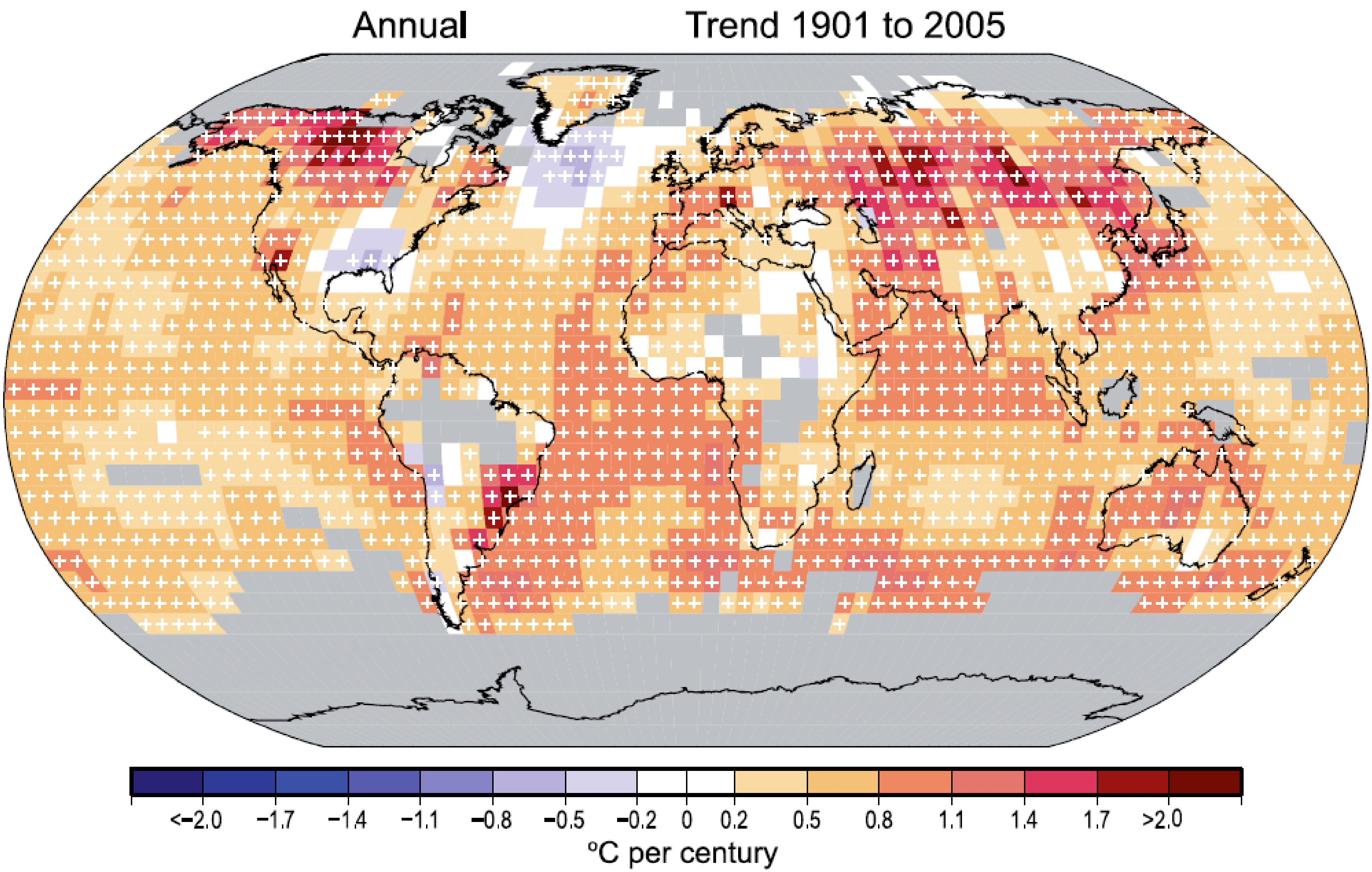
Schneider 2001



Loss of spatial information

Source: NASA GISS

Local linear temperature trends



Loss of temporal information

Source: IPCC 2007

Space-time filtering

- Anthropogenic climate change occurs on timescales of decades or longer
- To identify anthropogenic climate changes, isolate slow manifold of climate variations
- Use spatial correlations to devise more efficient low-pass filter than is obtainable from local information alone

Isolate slow climate variations with spatial and temporal information

Slow subspace of climate variations

- Define slow and fast covariance matrices $\Sigma_{>}$ and $\Sigma_{<}$
- Seek linear combinations $y = u^T x$ that maximize generalized Rayleigh quotient

$$\mathcal{R} = \frac{u^T \Sigma_{>} u}{u^T \Sigma_{<} u}$$

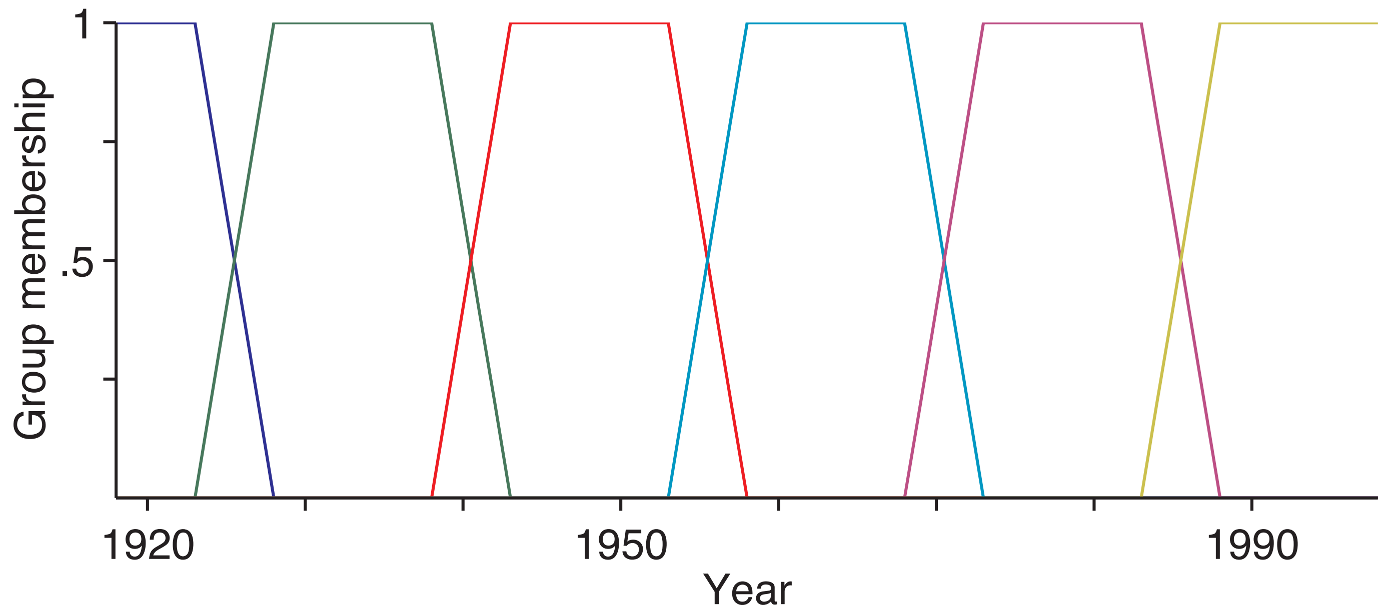
- Seek next linear combination $y = u^T x$ that maximizes $\mathcal{R}$ subject to being uncorrelated with first, etc.

Decomposition of variations into uncorrelated subspaces with decreasing ratio of slow to fast variance

Discriminant analysis

- Maximizes ratio of among-group to within group variance $\mathcal{R} = u^T \Sigma_{>} u / (u^T \Sigma_{<} u)$
- Groups are years of data (with fractional membership):

G



- Leads to generalized eigenvalue problem $\Sigma_{>} u = \gamma \Sigma_{<} u$

Schneider & Held 2001

Discriminant analysis (cont.)

- $\Sigma_<$ (or total covariance matrix Σ) is rank deficient
- Regularization by truncated PCA of Σ with effective rank chosen by GCV of regression $G = XA + \varepsilon$
- Results in weight vectors u and time series $y = u^T x$
- Spatial patterns v associated with time series y obtained by regression of x on y (dual of u in full rank case)
- Truncation of discriminant analysis to variates with $\mathcal{R} > 1$ (bootstrap) yields slow subspace of dimension r:

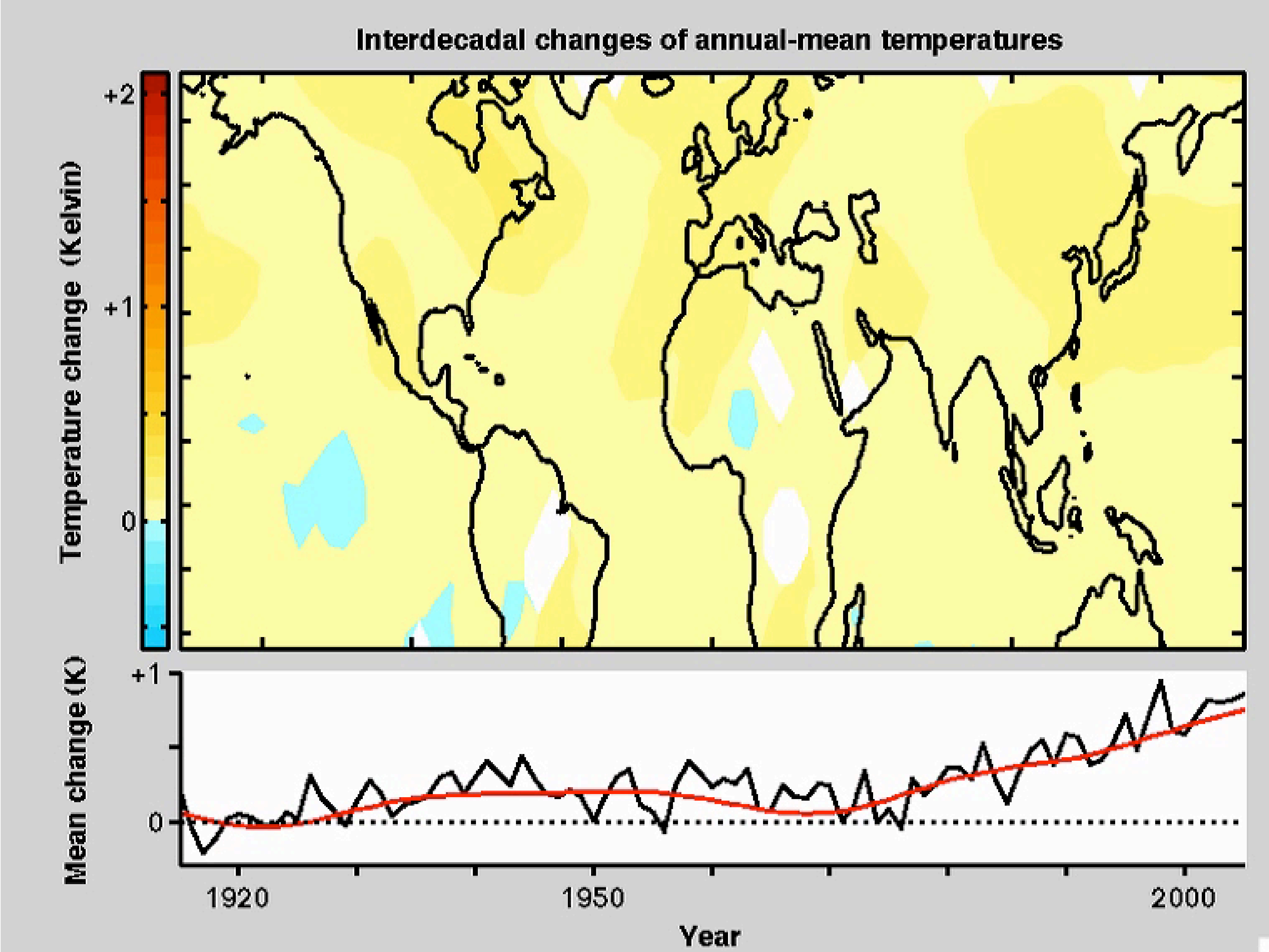
$$X_> \approx \sum_{i=1}^{r} y_i v_i^T$$

Schneider & Held 2001

Interdecadal temperature variations

- HadCRUT2 surface temperature data for 22.5S to 67.5N for years 1915 to 2005
- Eliminate grid points with more than 70% missing values
- Impute missing values and estimate covariances with regularized EM algorithm for remaining points
- Perform discriminant analysis to isolate interdecadal variations

Yields three-dimensional slow subspace



Schneider & Held, 2001; http://www.gps.caltech.edu/~tapio/discriminants/animations.html

Applications

- For climate change detection, evaluate similarity of slow manifolds of simulations and observations

- For model evaluation, evaluate differences between slow manifolds of simulations and observations

- Focus on slow manifolds may eliminate need for large ensembles because typically only the slow manifold is of interest in climate studies

Conclusions

- Regularized EM algorithm provides framework for estimation of missing values and covariance matrices in incomplete, rank-deficient data
- Different regularization approaches can be used within regularized EM algorithm (and should be tested):
  - Tikhonov/ridge regression
  - Truncated TTLS
  - Tapered covariance functions etc.
- Convergence of algorithm assured with fixed regularization parameter, but adaptive regularization parameter desirable (choose by GCV, GML, etc.)

Conclusions (cont.)

- Space-time filtering much more effective than local filtering to isolate slow variations
- Approach derived from discriminant analysis can be used to identify slow subspace of climate variations
- May be improved by allowing more flexible variance structures and by relaxing restriction to linear subspaces
- Effective space-time filtering may make large ensembles of climate simulations unnecessary