# Dimension Reduction and Sampling

Imola K. Fodor and Chandrika Kamath

Center for Applied Scientific Computing
Lawrence Livermore National Laboratory

SciDAC All-Hands Meeting, San Diego
September 11-13, 2002

UCRL-PRES-149962. This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract W-7405-Eng-48.

# We are investigating dimension reduction and sampling techniques

- Problem: data from simulations and experiments is high dimensional (i.e. many features)
- Querying the features can help in understanding the data
  - but, searching in a high-dimensional space is difficult
- May want to cluster similar objects for efficient access
  - but, clustering is expensive in high dimensions
- May want to analyze data
  - a representation in fewer dimensions would help
- Solution: use dimension reduction techniques
- But, dimension reduction techniques can be expensive if have many data items
- Solution: use sampling to appropriately reduce the number of data items

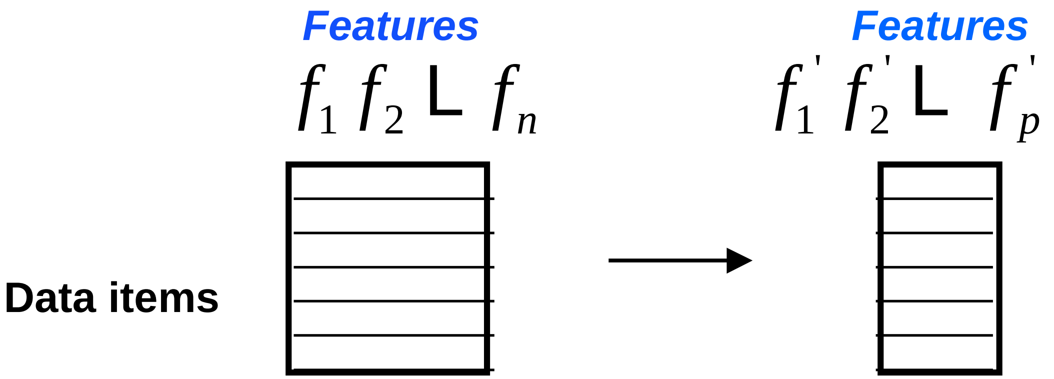
# Our work on dimension reduction will help both data management and mining

- **Reducing the dimensions will improve**
  - **searching (LBNL)**
  - **clustering (ORNL)**
- **Dimension reduction can also help in data mining and scientific discovery ➔ focus of this talk**
- **Our initial focus is on climate data**
  - **complements work at ORNL on climate**
- **Our techniques are also applicable to other data**
  - **high-energy-physics data LBNL on HEP**

**➔ We only discuss the .8 FTE work funded under SciDAC; however, our data mining research is more extensive. See www.llnl.gov/casc/sapphire**

# There are two different ways in which we can view dimension reduction

- **Reduce the number of features representing a data item**



$$p < n$$

- **Reduce the number of basis vectors used to describe the data: if some of the $\alpha_{ij}$ are small, they can be ignored**

$$DataItem_i = \sum_{j=1}^{N} \alpha_{ij}\ BasisVector_j$$

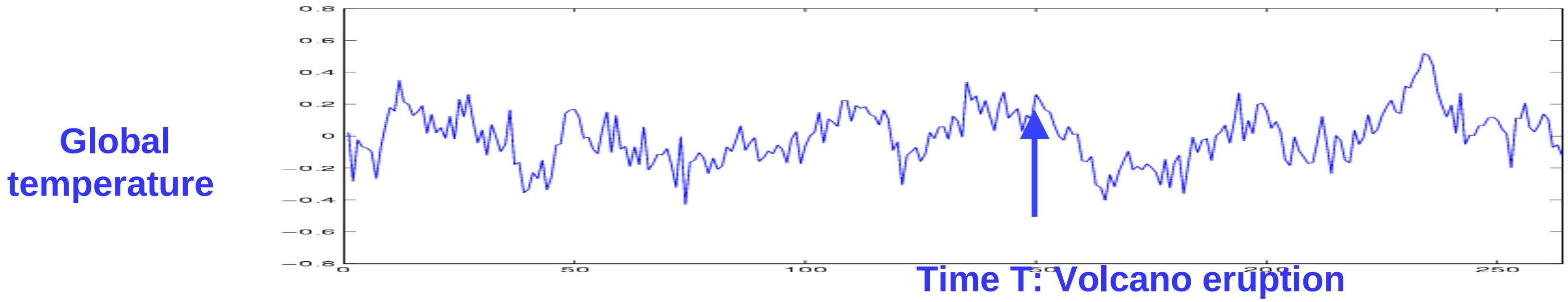
➔ **Dimension reduction can find a reduced representation**

# Our work on climate data focuses on reducing the number of basis vectors

- Atmospheric scientists are interested in understanding changes in global temperatures
- Simulated and observed data include effects of volcano eruptions, El Niño and Southern Oscillation (ENSO), etc.
- We need to remove effects that are not shared by the different models to
  - —make meaningful comparisons
  - —understand effects of man-made contributions for global warming
- Domain expert Dr. Benjamin Santer (PCMDI, LLNL)
  - —MacArthur award for research supporting the finding that human activity contributes to global warming

➔ Dimension reduction supporting scientific discovery

# Isolating the effects of different sources is a difficult problem



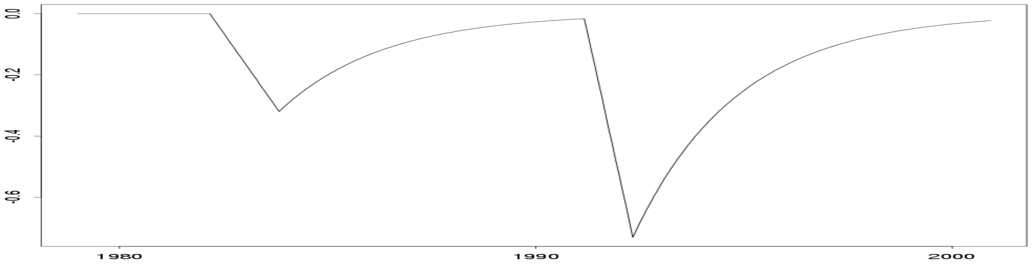
How much of the cooling after time T is volcano induced?

- **Separation is difficult as El Chichón and Mt. Pinatubo volcano eruptions coincided with ENSO events**
- **Traditional methods such as principal components (PCA) on the global mean series have not been successful**
- **Current approaches don't always work**
- **Need better understanding of the**
  - **interaction between signals**
  - **conditions under which methods work, and why**

➔ **Active area of research in climate**

# Current techniques for separating volcano and ENSO signals use parametric models

- Best current approach
  - —create parametric models for volcano and ENSO signals
  - —estimate and remove ENSO effect
  - —estimate and remove volcano effect
  - —iterate



$$v_t = \begin{cases} \dfrac{-\Delta T_m t}{t_{ramp}}, & t = t_{erupt}, \ldots, t_{ramp} \\ -\Delta T_m e^{-\frac{t - t_{ramp}}{\tau}}, & t = t_{ramp}+1, \ldots, T \end{cases}$$

**A model for the effects of two volcano eruptions on global temperatures:**

**Known parameters:** $T = 264; \ \tau = 30; \qquad t^1_{erupt} = 39, \ t^2_{erupt} = 147;$

**Est. parameters:** $\Delta T^1_m = 0.32; \quad t^1_{ramp} = 20 \qquad \Delta T^2_m = 0.72; \quad t^2_{ramp} = 14$

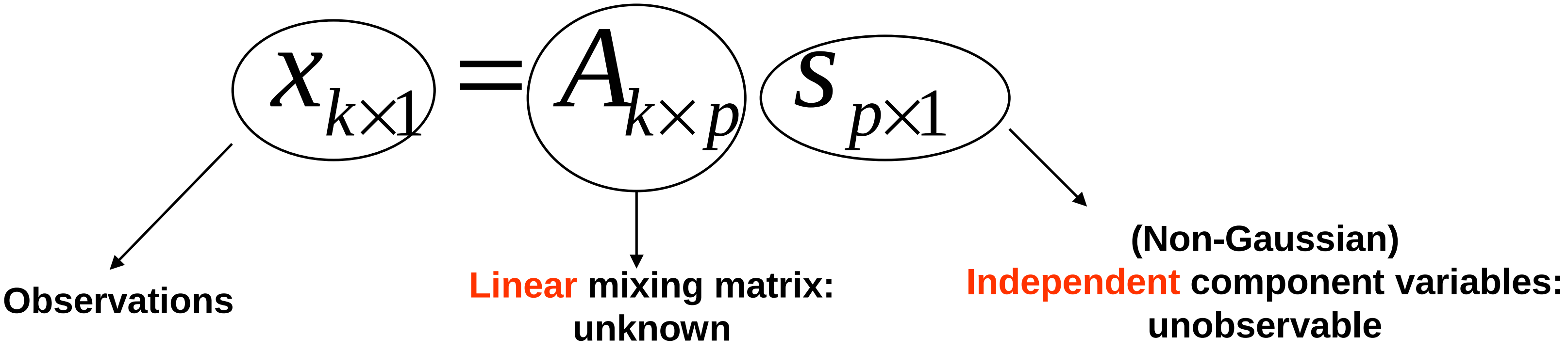
# To complement the parametric models, we investigate automated techniques

- Parametric approach [1] has many drawbacks
  - different estimation techniques lead to different parameter estimates
  - it is sensitive to parameter values: slightly different parameters lead to different results
  - what if signals do not follow the proposed models?
- Can automated techniques help?
  - use the data itself to drive the separation of signals
  - explore independent component analysis (ICA)
- Can zonal signals give better results than global signals?

[1] B.D. Santer et al. Accounting for the effects of volcanoes and ENSO in comparisons of modeled and observed temperature trends. *J. Geophys. Res*. 106, D22, Nov. 27, p. 28,033--28,059, 2001.

# ICA assumes that the observations are linear mixtures of unobservable variables

- Simplest ICA model



$$x_{k\times 1} = A_{k\times p}\, s_{p\times 1}$$

Observations

Linear mixing matrix: unknown

(Non-Gaussian) Independent component variables: unobservable

- Given n realizations of x, estimate A and s
- Connection to PCA [6]
  - for Gaussian variables, ICA = PCA
  - PCs are uncorrelated, while ICs are independent
- ICA is very active research area, new developments, extensions to more complicated models are currently under investigation [2,3,4,5]

# ICA seeks independent components by optimizing measures of independence

- **E.g. minimize the mutual information**

  $$I(y) = J(y) - \sum_{i=1}^{n} J(y_i)$$

  **for the uncorrelated** $y = (y_1, \ldots, y_n)$ **with joint probability density function** $f(y)$**, where**

  $J(y)$ **is the negentropy:** $J(y) = H(y_{gauss}) - H(y)$

  $H(y)$ **is the entropy:** $H(y) = -\int f(y) \log f(y)\, dy$

  **and** $y_{gauss}$ **is Gaussian s.t.** $Cov(y_{gauss}) = Cov(y)$

- **Various approximations and computational tricks**

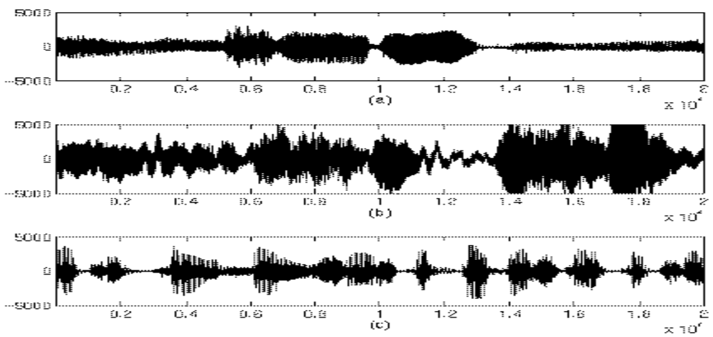
  $$J(y_i) \approx [E\{G(y_i)\} - E\{G(\upsilon)\}]^2$$

  **where** $\upsilon \approx N(0,1)$**, and** $G(.)$ **is a suitable non-quadratic function, such as** $G(u) = \log\cosh(u)$

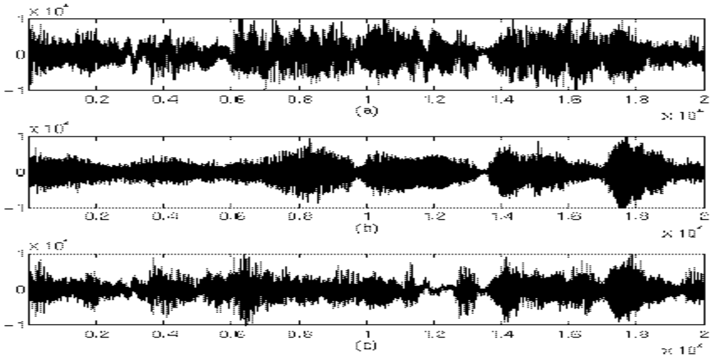
- ***fastICA* software from http://www.cis.hut.fi/~aapo**

# ICA separates individual signals from mikes that record simultaneous speakers

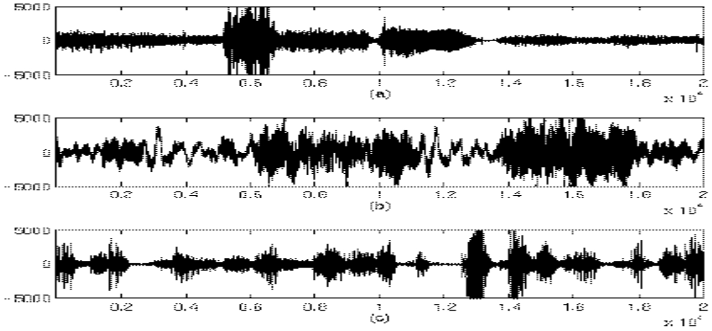
- The cocktail party problem: many online demos
  - http://www.mns.brain.riken.go.jp/~shiro/blindsep.html
  - http://www.cnl.salk.edu/~tewon/Blind/blind_audio.html
  - http://www-sigproc.eng.cam.ac.uk/oldusers/dcbc1/research/diagram.html



(i) 3 sources



(ii) 3 observations



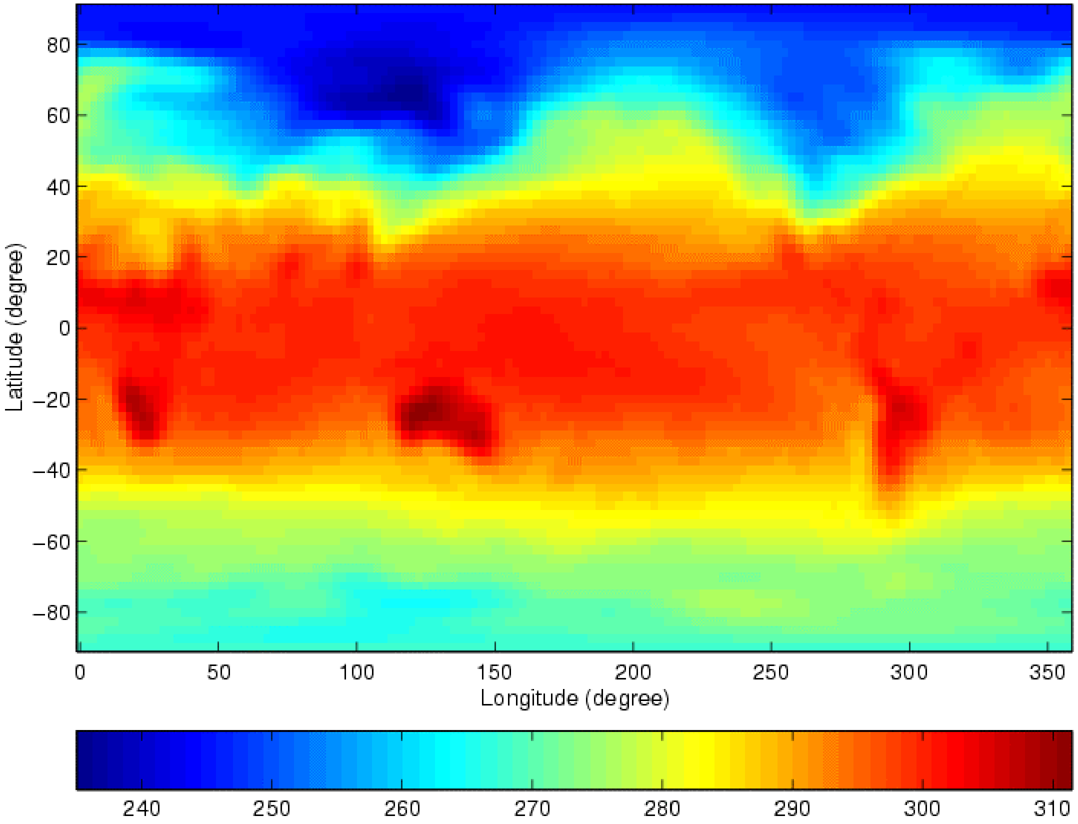
(iii) 3 estimated sources from (ii) after ICA

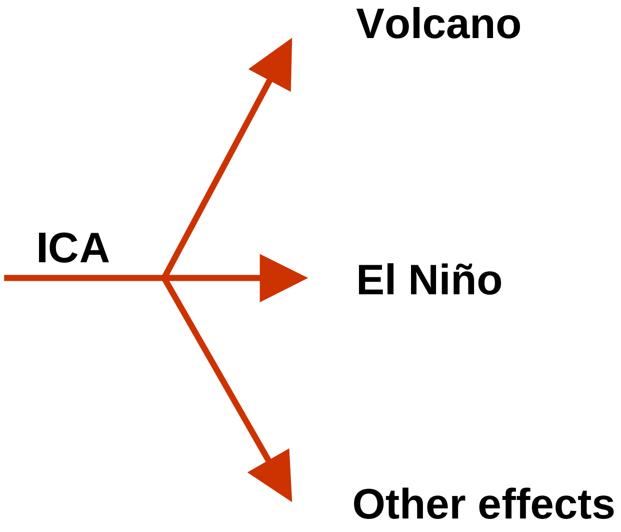
# ICA has also been successfully applied in other source separation problems

- **Removing artifacts from EEG/MEG brain data**
  - **Measure brain activity on the scalp by removing un-related artifacts, such as eye-blinks**
- **Removing train signals from seismograms**
  - **Study earthquake activity by isolating train noise from seismograms**
- **Economic time series, telecommunications, …, [2,3,4,5]**
- **The similarities with our climate problem prompted us to investigate ICA in our context**

**➔ To our knowledge, ours is the first attempt to consider ICA in the atmospheric sciences**

# The raw data: 264 monthly temperatures on a 144x73 spatial grid on 17 vertical levels




ICA

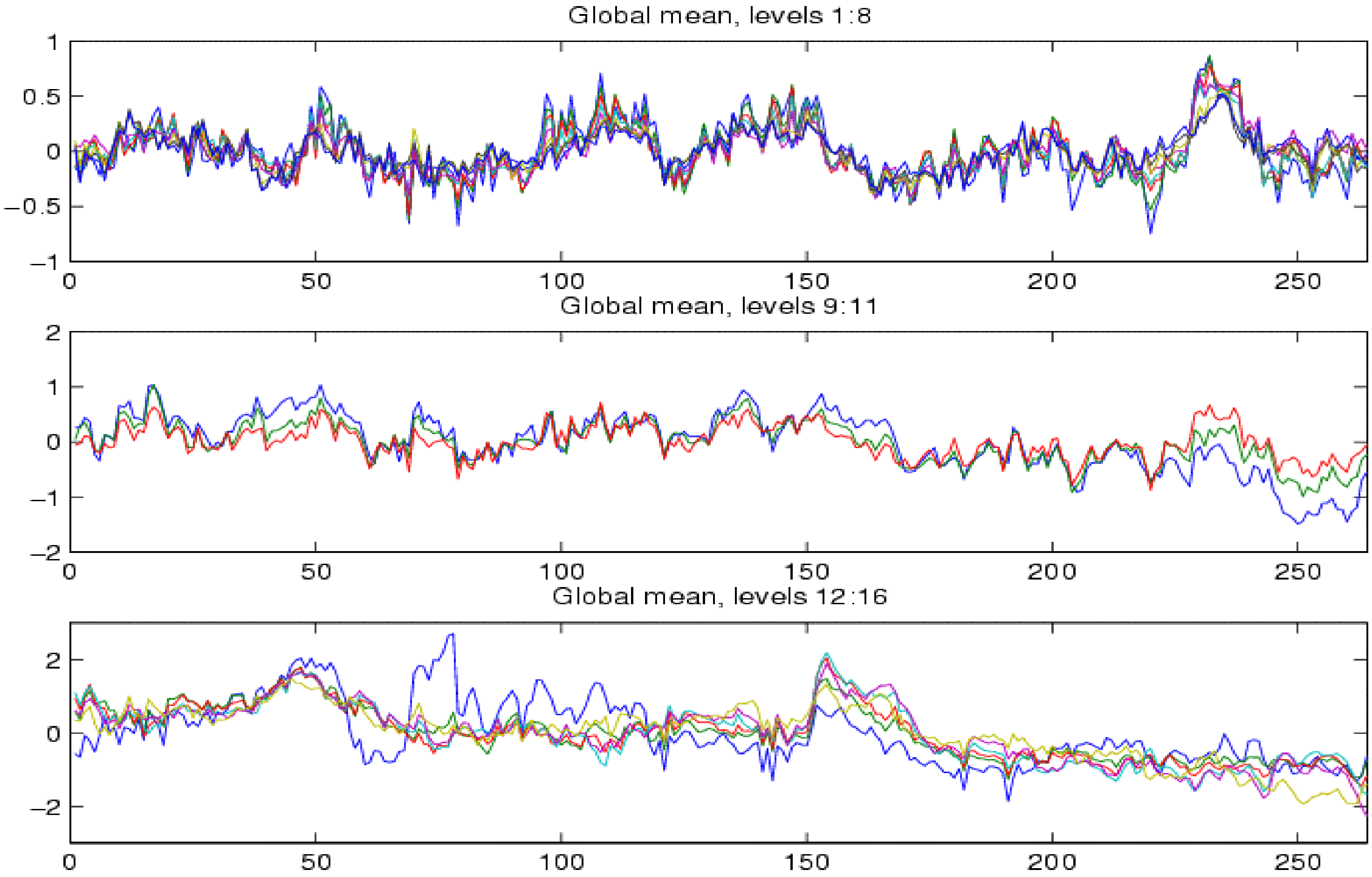
Volcano

El Niño

Other effects

**January 1979 raw temperatures (Kelvin) on the 144x73 latitude by longitude grid at 1000hPa pressure level. Data from NCEP.**

# Climate scientists typically work with global monthly means data



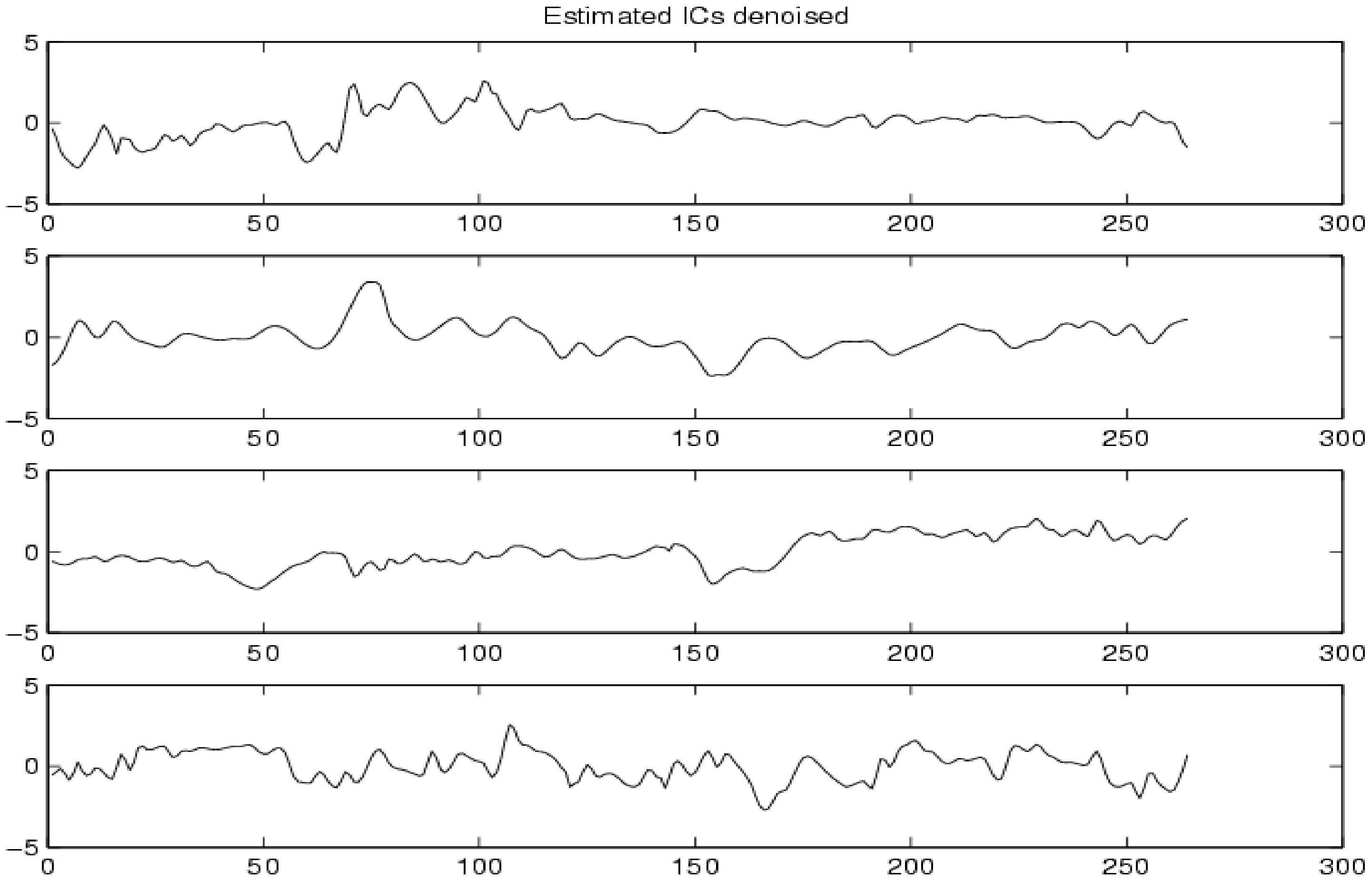
**17 vertical levels**

**level 1: 1000hPa, lowest altitude**

**level 17: 10hPa, highest altitude**

**Time series of global monthly mean anomalies, Jan 1979 - Dec 2000**

# IC estimates (denoised) based on global temperatures from the four lowest levels



???

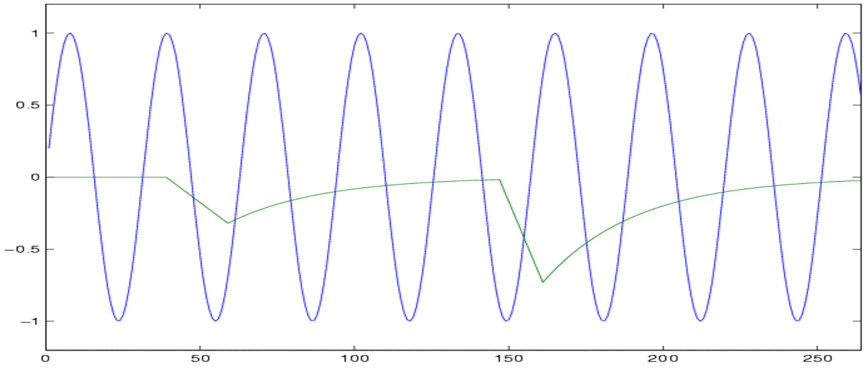
???

Volcano signal?

???

➔ **Difficult to interpret the estimates: use synthetic data**

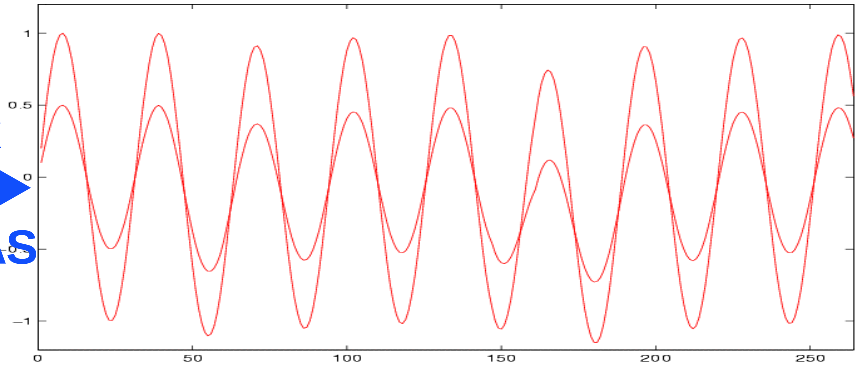
# We experimented with synthetic data to understand the behavior of ICA



(i) Two IC sources: sine $S_1$ and volcano $S_2$

Mix

X=AS

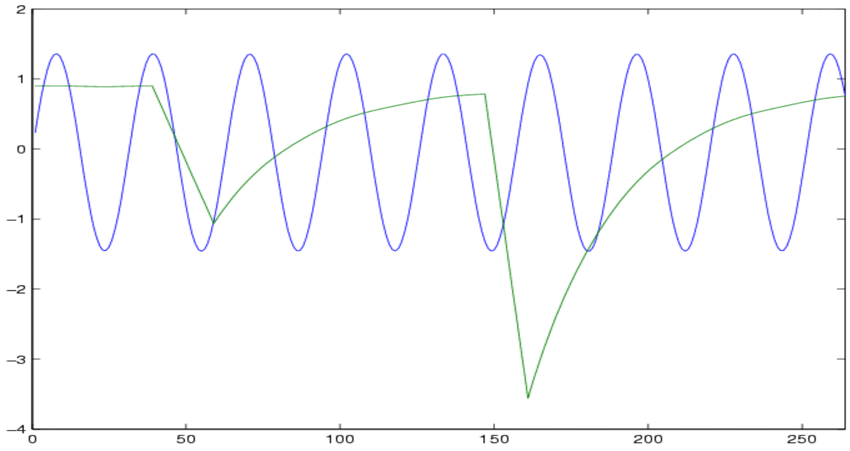


(ii) Two mixed signals: $X_1$ and $X_2$

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \underbrace{\begin{pmatrix} 1.0 & 0.4 \\ 0.5 & 0.6 \end{pmatrix}}_{} \begin{pmatrix} S_1 \\ S_2 \end{pmatrix}$$

Mixing matrix: $A$

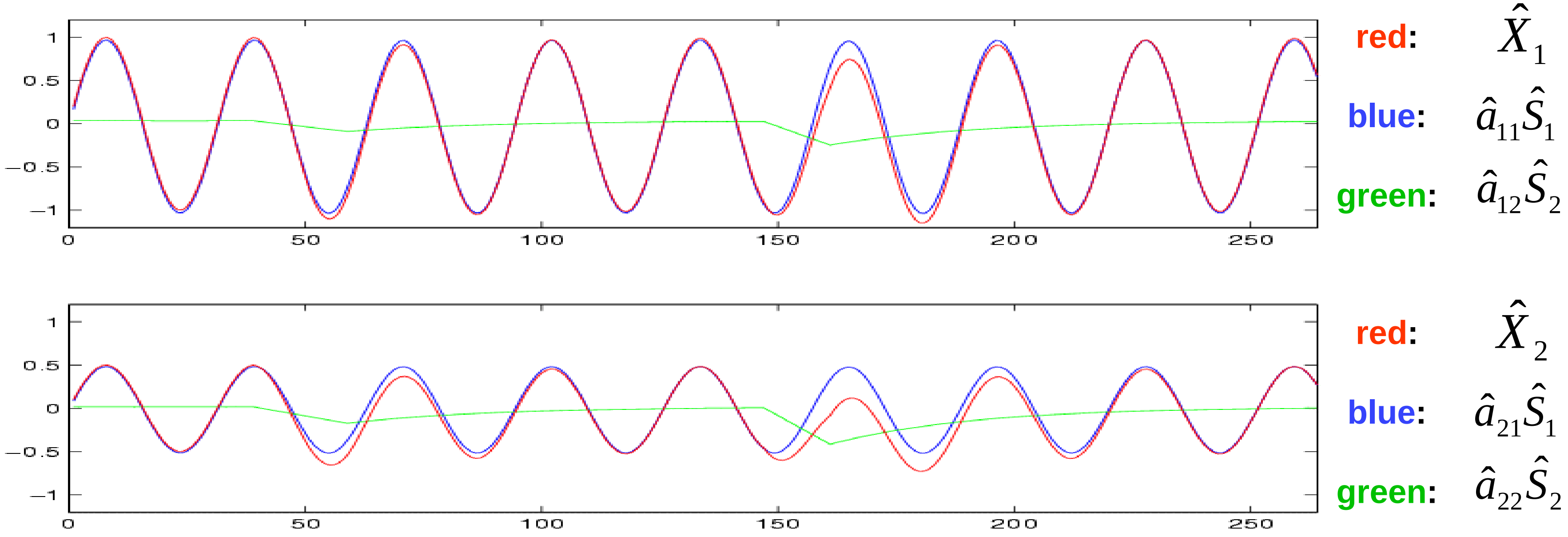
ICA



(iii) Sources estimated from (ii): $\hat{S}_1$ and $\hat{S}_2$

The fastICA algorithm estimates correctly the shapes of the two independent components (ICs), but not their respective amplitudes.

# With proper post-processing, we can also estimate accurately the IC amplitudes

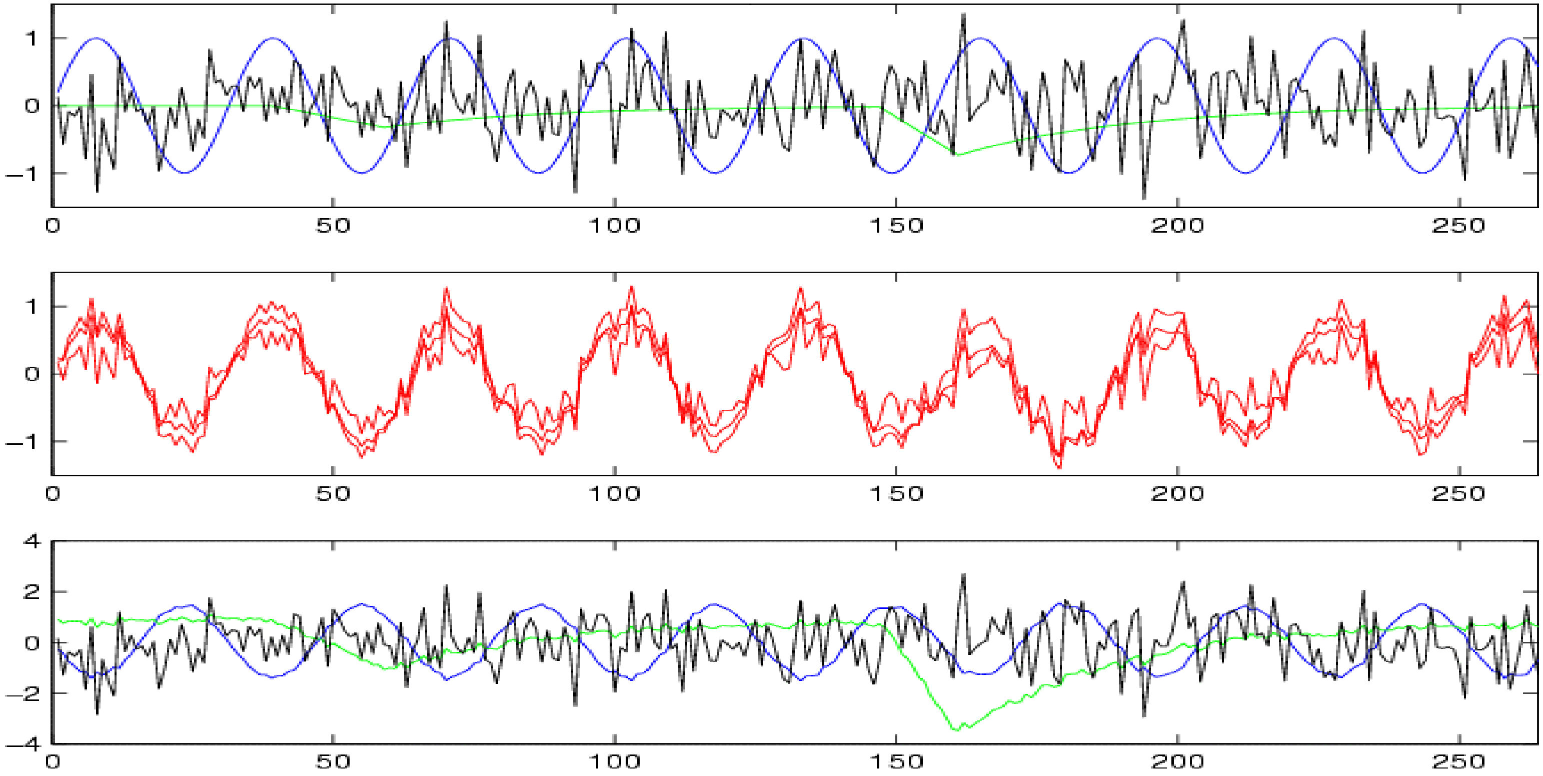


**The mixed signals in terms of the estimated independent components**

$$\begin{pmatrix} \hat{X}_1 \\ \hat{X}_2 \end{pmatrix} = \hat{A} \begin{pmatrix} \hat{S}_1 \\ \hat{S}_2 \end{pmatrix} = \begin{pmatrix} \hat{a}_{11} & \hat{a}_{12} \\ \hat{a}_{21} & \hat{a}_{22} \end{pmatrix} \begin{pmatrix} \hat{S}_1 \\ \hat{S}_2 \end{pmatrix}$$

➔ **Ben Santer: the automatic separation is “very impressive”**

# Since most scientific data is noisy, we explored the robustness of ICA to noise



**3 IC sources:**

**sine**: $S_1$

**volcano**: $S_2$

**noise**: $S_3$

**Mix** ↓ **X=AS**

**3 mixed signals:**

$X_1, X_2, X_3$

↓ **ICA**

**3 IC estimates:**

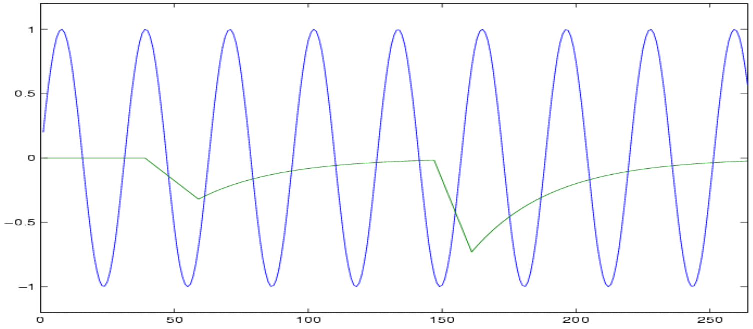
**sine**: $\hat{S}_1$

**volcano**: $\hat{S}_2$

**noise**: $\hat{S}_3$

➔ **ICA can separate noise used as an extra component**

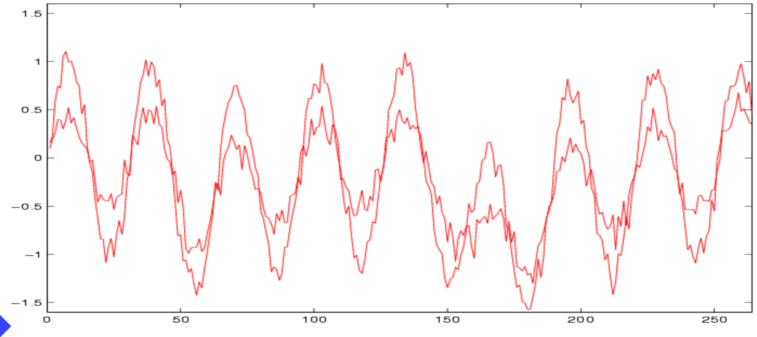
# ICA, combined with wavelet denoising, is fairly robust to noise added after mixing



(i) Two IC sources: sine $S_1$ and volcano $S_2$

Mix, then

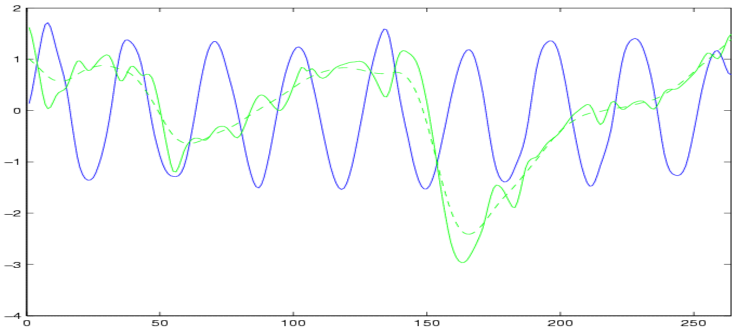
add noise



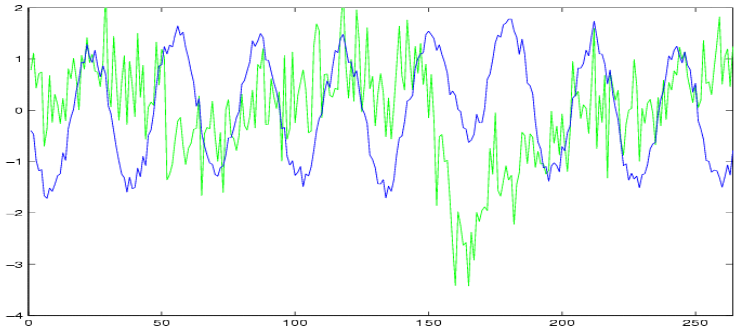
(ii) Two mixed signals + noise: $Y_1$ and $Y_2$

ICA

Denoise, then ICA



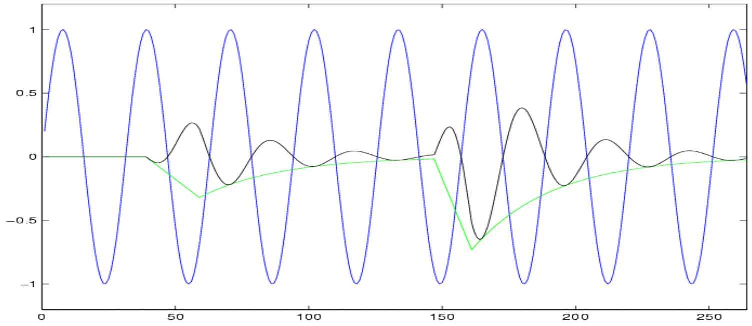
(iii) Sources estimated from (ii) by first denoising it, then using ICA



(iv) Sources estimated from (ii)

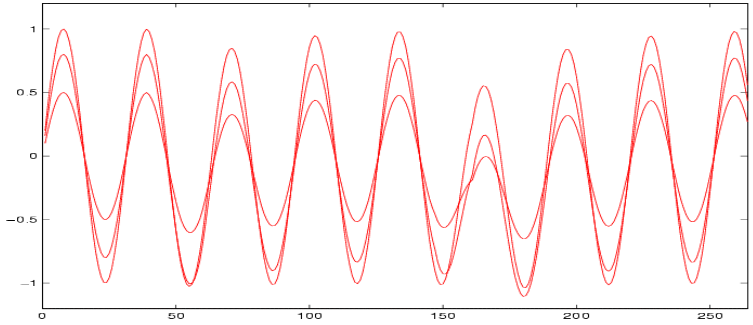
$$\begin{pmatrix}Y_1\\Y_2\end{pmatrix} = A\begin{pmatrix}S_1\\S_2\end{pmatrix}+\begin{pmatrix}\varepsilon_1\\\varepsilon_2\end{pmatrix}$$

# We also explored the robustness of ICA to the independence assumption



Mix

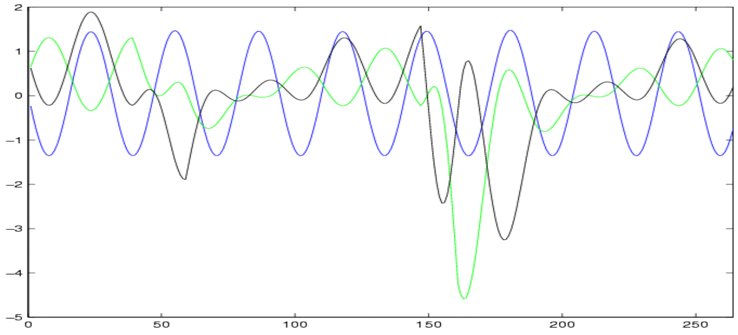
X=AS



(i) Three sources: sine $S_1$ , volcano $S_2$, and their interaction $S_3 = S_1 S_2$

(ii) Three mixed signals: $X_1, X_2$ and $X_3$

ICA

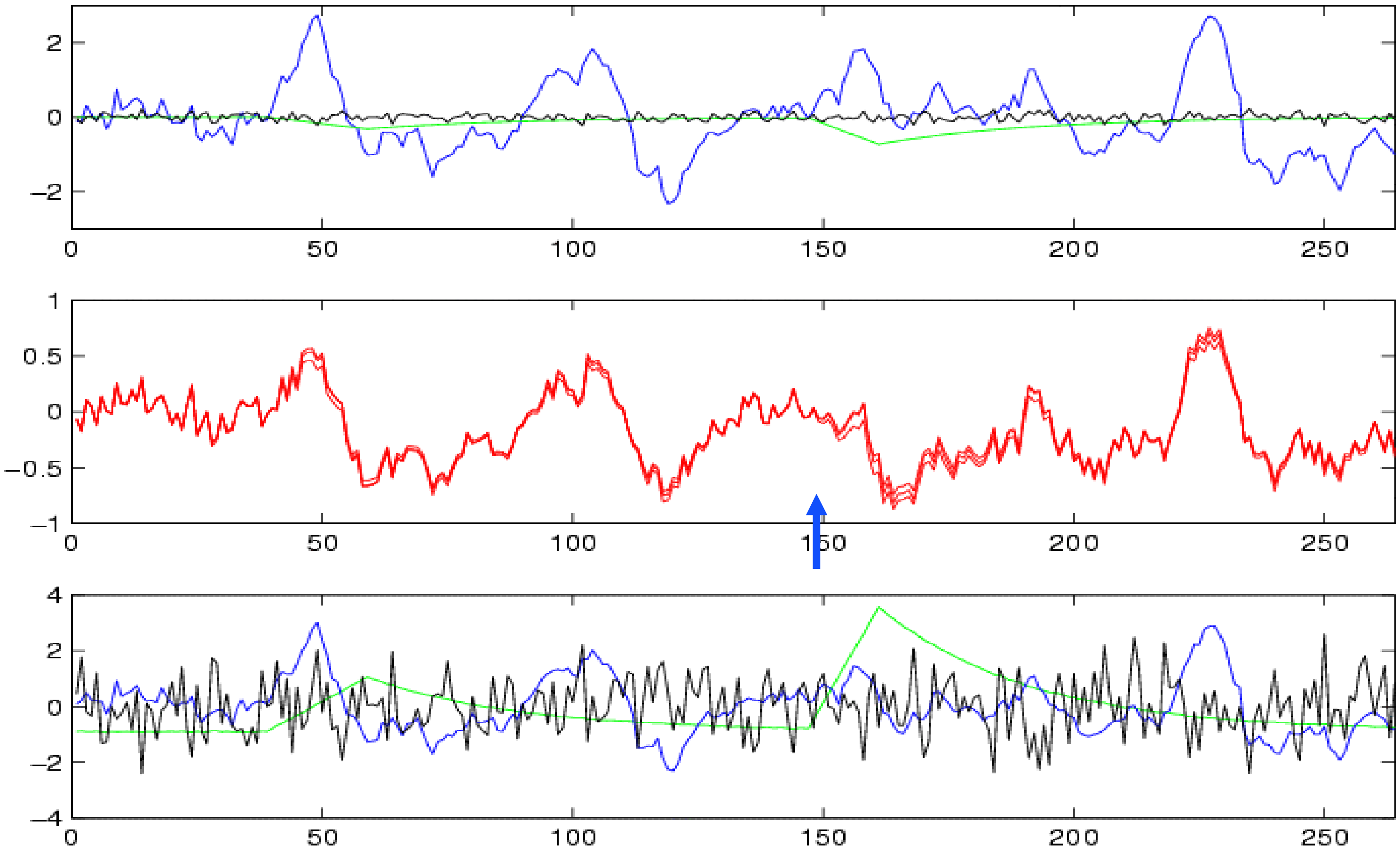


The simple ICA model cannot separate non-independent sources.

➔ **Ben Santer:**
**“negative result is valuable”**

(iii) Sources estimated from (ii)

# A more realistic model: three mixed signals = volcano + noise + El Niño (instead of sine)



**3 IC sources:**

**El Nino**: $S_1$
**volcano**: $S_2$
**noise**: $S_3$

**Mix** ↓ **X=AS**

**3 mixed signals:**
$X_1, X_2, X_3$

↓ **ICA**

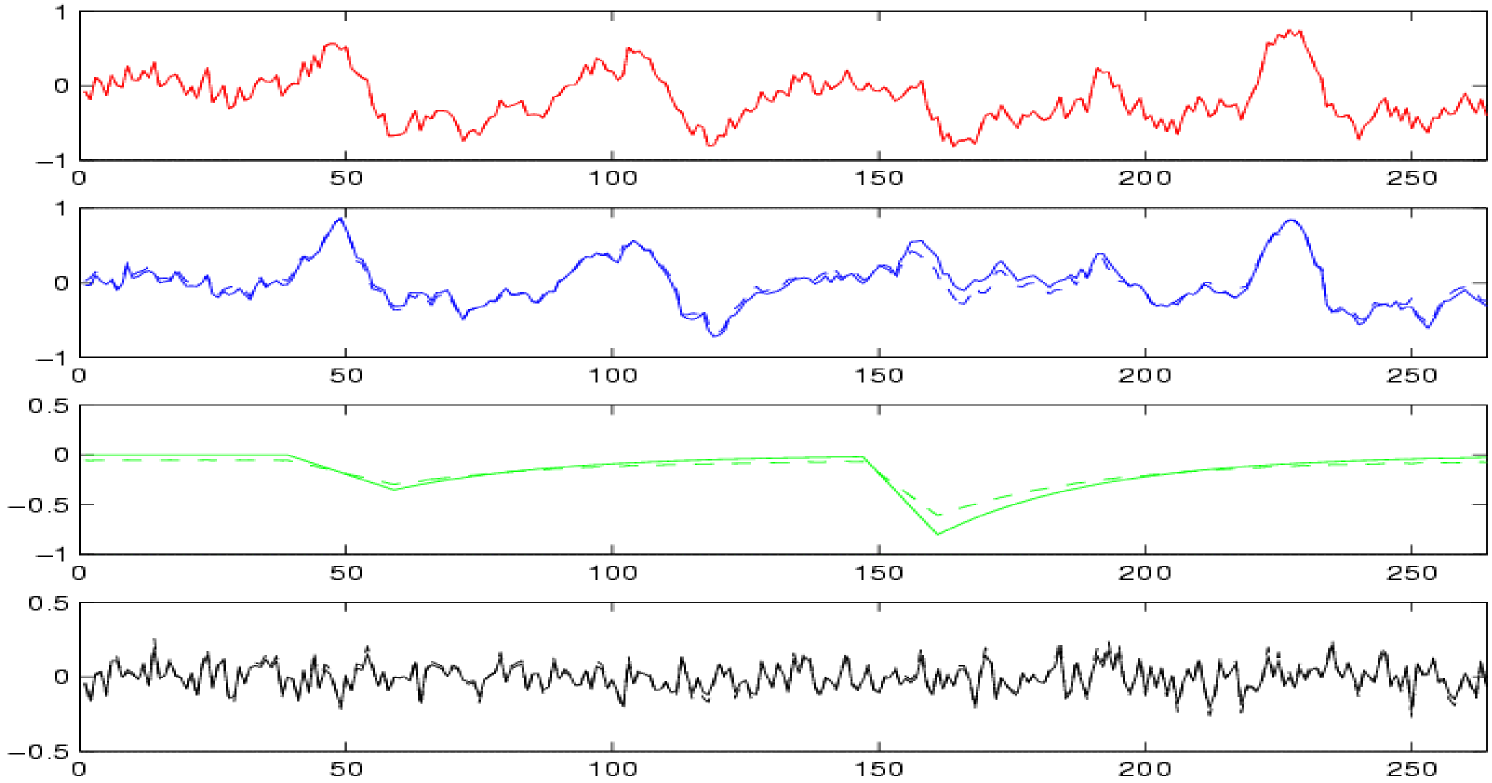
**3 IC estimates:**

**El Nino**: $\hat{S}_1$
**volcano**: $\hat{S}_2$
**noise**: $\hat{S}_3$

**Cooling in the mixed global signals after the arrow is in fact a combination of an El Nino warming and a volcano cooling. Without the volcano eruption, the global temperatures would be higher in this model.**

# The IC estimates are in excellent agreement with the known sources

- **Continuous lines represent the true decompositions, while dashed ones the ICA estimates**



**red:** $X_1$

**red dashed:** $\hat{X}_1$

**blue:** $a_{11}S_1$

**blue dashed:** $\hat{a}_{11}\hat{S}_1$

**green:** $a_{12}S_2$

**green dashed:** $\hat{a}_{12}\hat{S}_2$

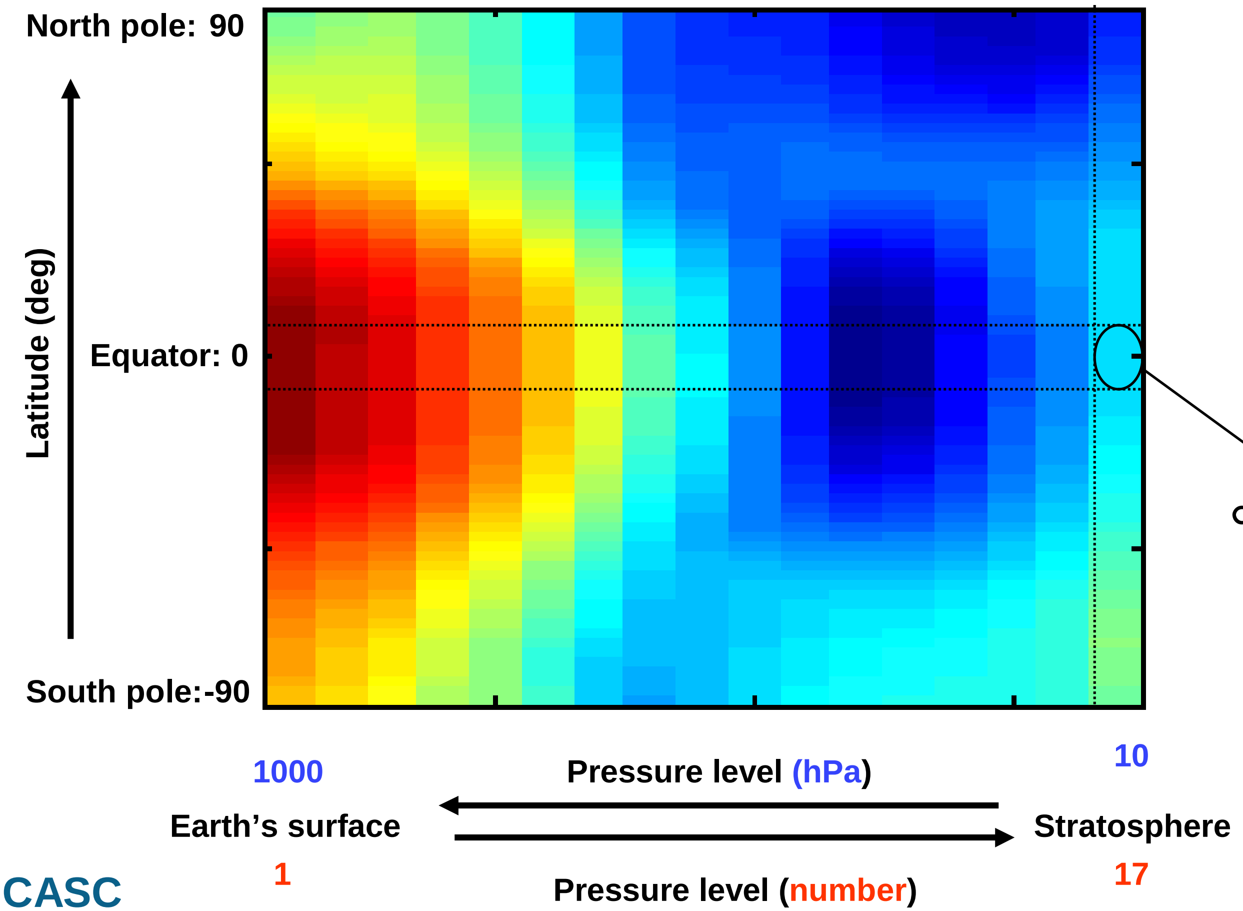
**black:** $a_{12}S_2$

**black dashed:** $\hat{a}_{12}\hat{S}_2$

$$X_1 = a_{11}S_1 + a_{12}S_2 + a_{13}S_3 \qquad \hat{X}_1 = \hat{a}_{11}\hat{S}_1 + \hat{a}_{12}\hat{S}_2 + \hat{a}_{13}\hat{S}_3$$

# Ben Santer suggested ICA on zonal data to search for spatial source signatures

- **Monthly means for 73 zones on 17 vertical levels: Jan 1979 – Dec 2000**



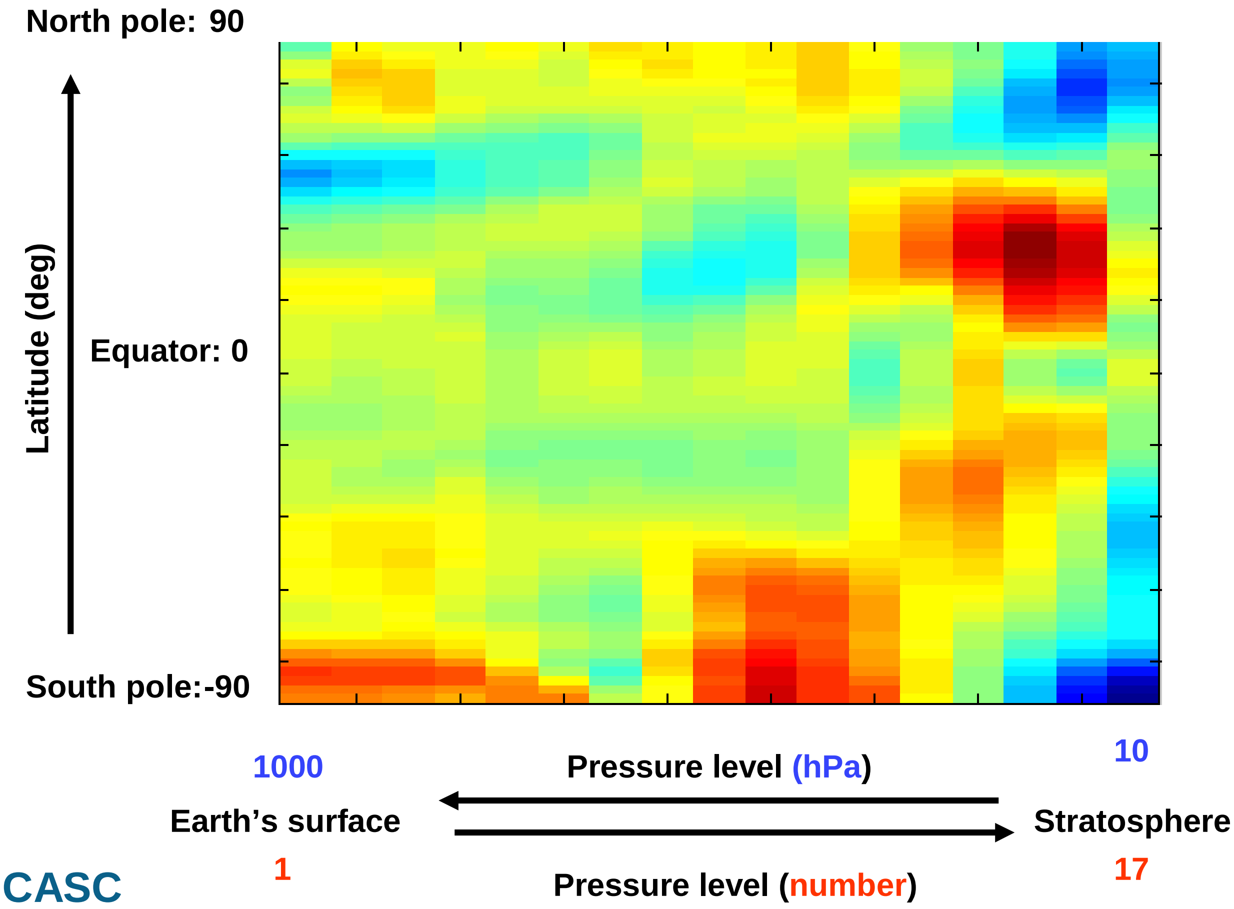
**Zone: a latitude band.**

**Zonal mean: average over all longitudes, on a given latitude band on a given vertical level.**

$$\propto \sum_{lat=-2.5}^{2.5} \sum_{lon=0}^{360} t_{lat,lon,lev=17}$$

# Zonal monthly mean anomaly data

- Monthly mean anomalies for 73 zones on 17 vertical levels: Jan 1979



North pole: 90

Latitude (deg)

Equator: 0

South pole:-90

1000

Pressure level (hPa)

10

Earth’s surface

Stratosphere

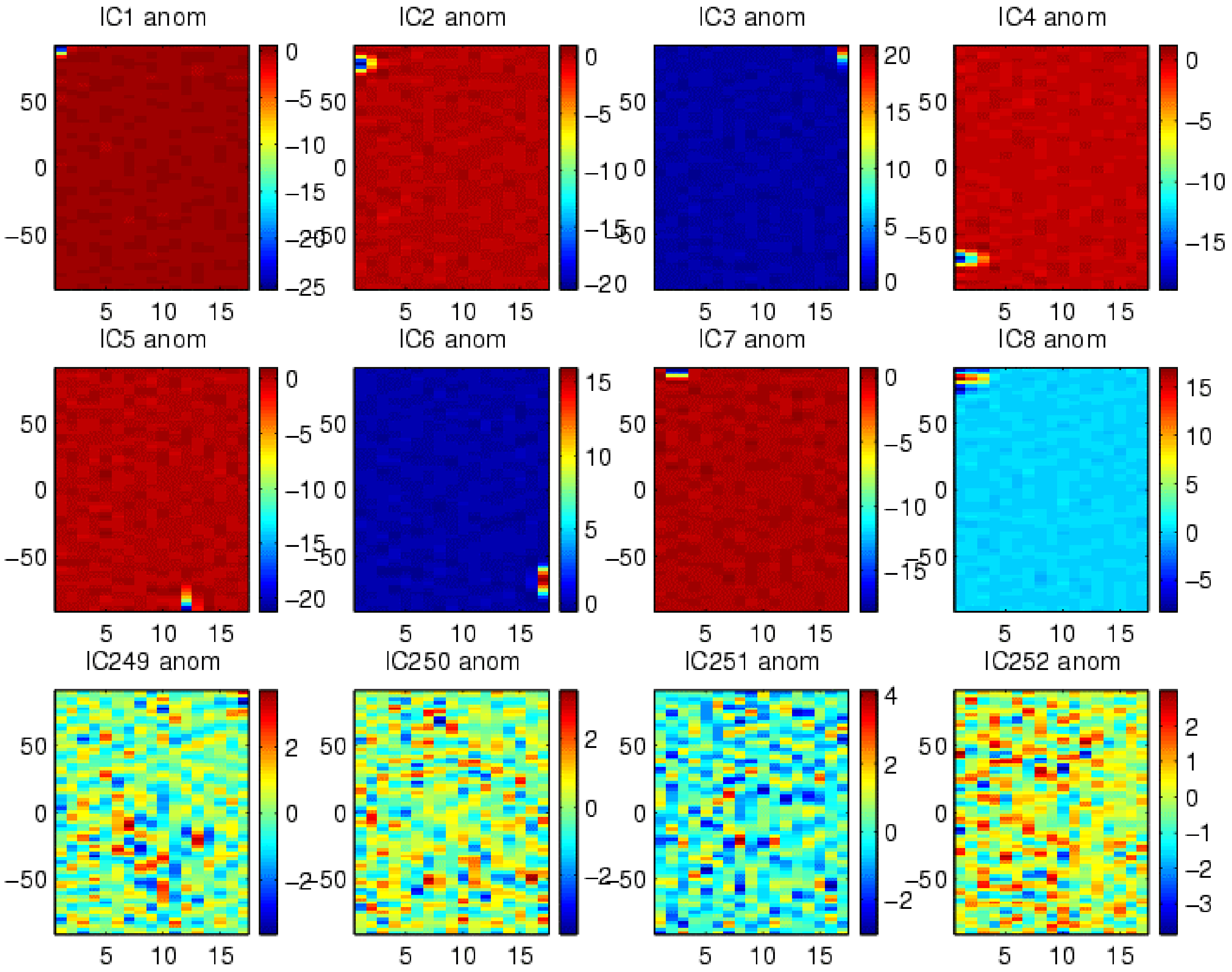
1

Pressure level (number)

17

Anomaly: departure from mean over 1979-2000.

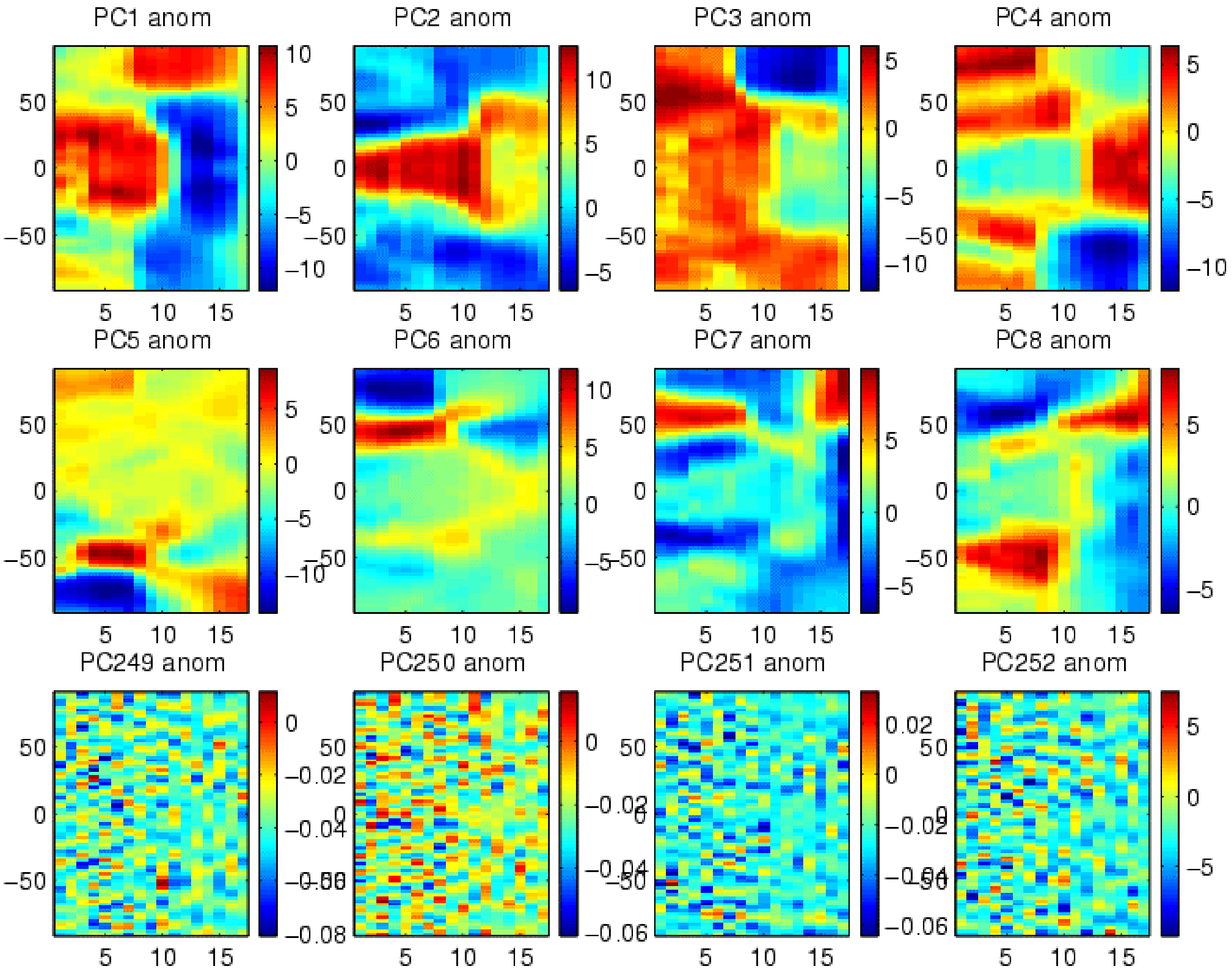
# Example ICs for the zonal anomaly data



**Not clear how to interpret the estimates. They are independent, but do not correspond to known physical phenomena.**

CASC

# Example PCs (cov matrix) for the zonal anomaly data



| #PC | Cumulative %Variation |
|---|---|
| 1 | .15 |
| 5 | .46 |
| 10 | .66 |
| 25 | .88 |
| 50 | .96 |
| 252 | 1 |

Interpretation much more straightforward. Ben Santer was very pleased when we showed him our results, and suggested further analyses.

# Summary

- ICA separates linearly mixed signals in
  - —synthetic data
  - —synthetic data with noise added
- ICA runs into problems with
  - —non-linear mixing of synthetic data
  - —real global means data => real data likely to be a non-linear mix of volcano and ENSO signals
- ICA results difficult to interpret if use zonal means instead of global means, but PCA appears promising
- Results presented at the Joint Statistical Meetings, Aug 2002, NYC

➔ Ben Santer: our work is helping him understand a new technique and its limitations in analyzing climate data

# References

- [1] B.D. Santer et al. Accounting for the effects of volcanoes and ENSO in comparisons of modeled and observed temperature trends. *J. Geophys. Res*. 106, D22, Nov. 27, p. 28,033--28,059, 2001.
- [2] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.
- [3] T.W. Lee. *Independent Component Analysis: Theory and Practice*. Kluwer, 2001.
- [4] S. Roberts and R. Everson, editors. *Independent Component Analysis: Principles and Practice*. Cambridge University Press, 2001.
- [5] M. Girolami, editor. *Advances in Independent Component Analysis*. Springer, 2000.
- [6] J. Friedman, T. Hastie, and R. Tibshirani. *Elements of Statistical Learning: Prediction, Inference and Data Mining*. Springer, 2001.
- [7] I.K. Fodor and C. Kamath. *On the use of ICA to separate meaningful sources in global temperature series*. In preparation.

# Dimension Reduction and Sampling: Mini Review

Imola K. Fodor and Chandrika Kamath

Center for Applied Scientific Computing
Lawrence Livermore National Laboratory

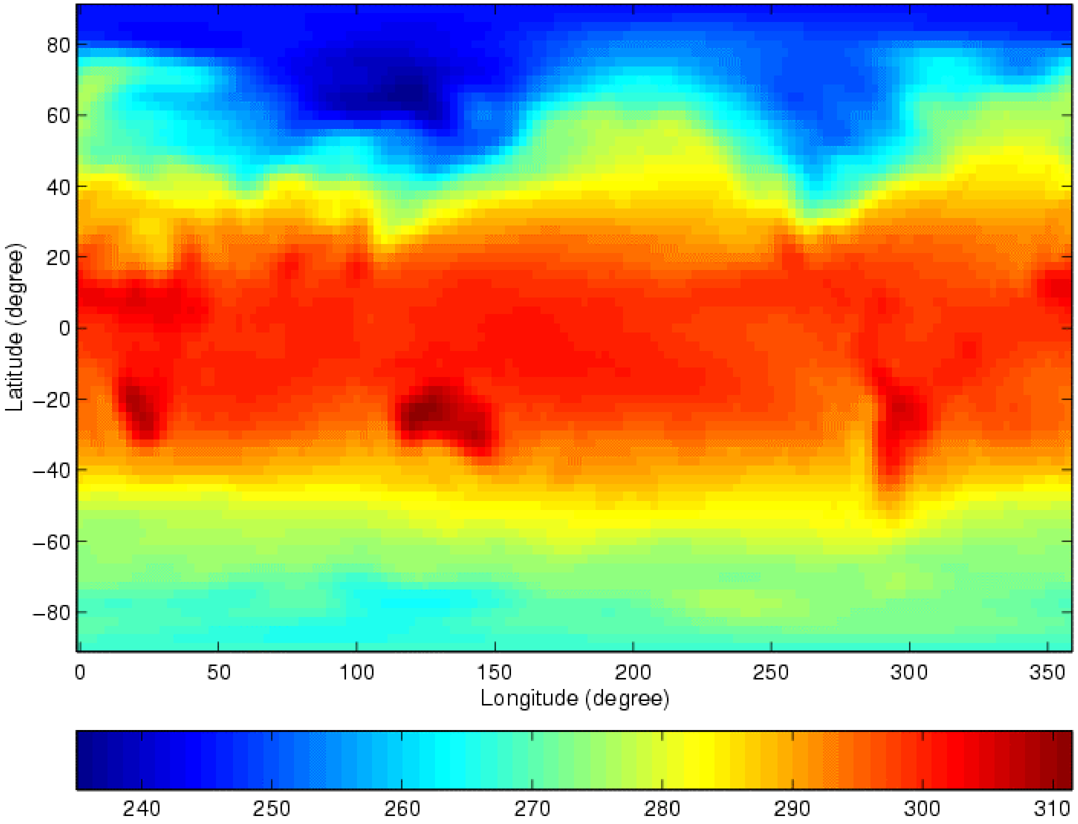
SciDAC All-Hands Meeting, San Diego
September 11-13, 2002

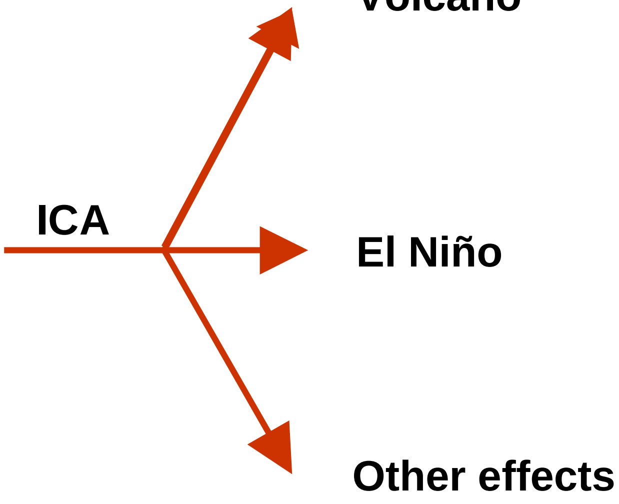
# Our work on climate data focuses on separating volcano and El Niño signals

- Atmospheric scientists are interested in understanding changes in global temperatures
- Simulated and observed data include effects of volcano eruptions, El Niño and Southern Oscillation (ENSO), etc.
- We need to remove effects that are not shared by the different models to
  - make meaningful comparisons
  - understand effects of man-made contributions for global warming
- Domain expert Dr. Benjamin Santer (PCMDI, LLNL)

➔ Dimension reduction supporting scientific discovery

# The raw data: 264 monthly temperatures on a 144x73 spatial grid on 17 vertical levels





ICA

Volcano

El Niño

Other effects

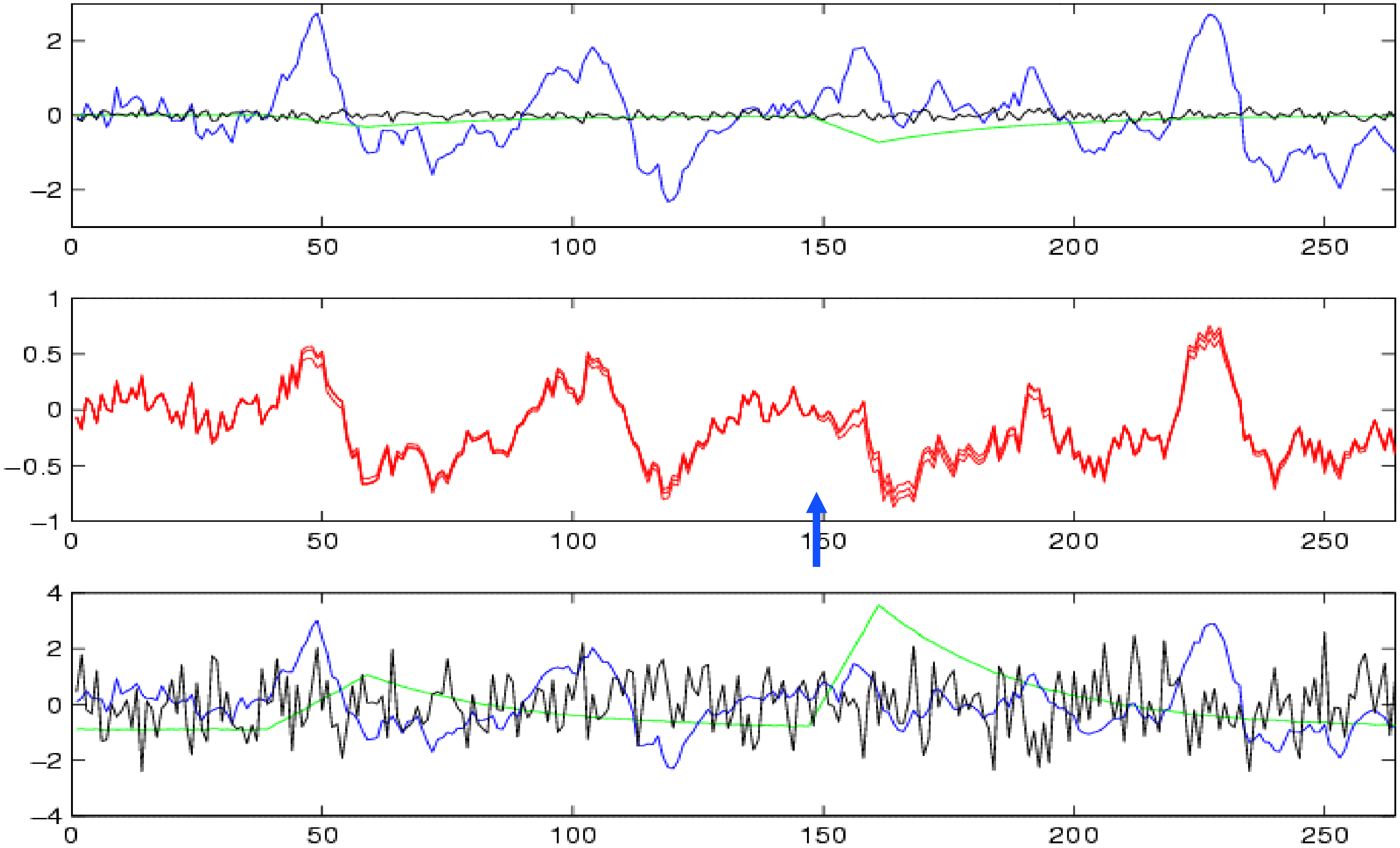
January 1979 raw temperatures (Kelvin) on the 144x73 latitude by longitude grid at 1000hPa pressure level.  Data from NCEP.

# Summary of work so far

- ICA separates linearly mixed signals in
  - synthetic data
  - synthetic data with noise added
- ICA runs into problems with
  - non-linear mixing of  synthetic data
  - real global means data => real data likely to be a non-linear mix of volcano and ENSO signals
- ICA results difficult to interpret if use zonal means instead of global means, but PCA appears promising
- Results presented at the Joint Statistical Meetings, Aug 2002, NYC

➔ Ben Santer: our work is helping him understand a new technique and its limitations in analyzing climate data

# A more realistic model: three mixed signals = volcano + noise + El Niño (instead of sine)



3 IC sources:

El Nino: $S_1$

volcano: $S_2$

noise: $S_3$

Mix X=AS

3 mixed signals:

$X_1, X_2, X_3$

ICA

3 IC estimates:

El Nino: $\hat{S}_1$

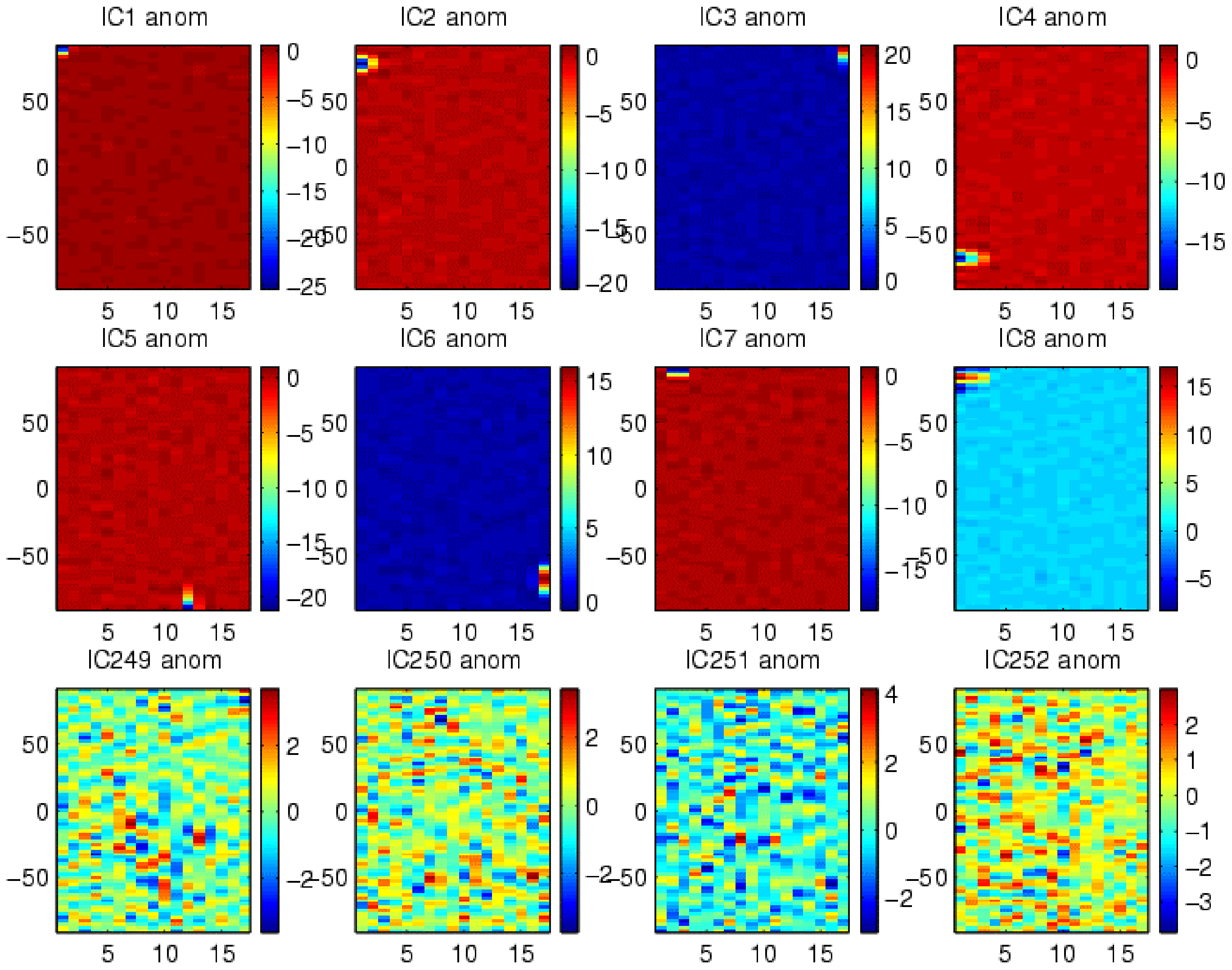
volcano: $\hat{S}_2$

noise: $\hat{S}_3$

Cooling in the mixed global signals after the arrow is in fact a combination of an El Nino warming and a volcano cooling. Without the volcano eruption, the global temperatures would be higher in this model.

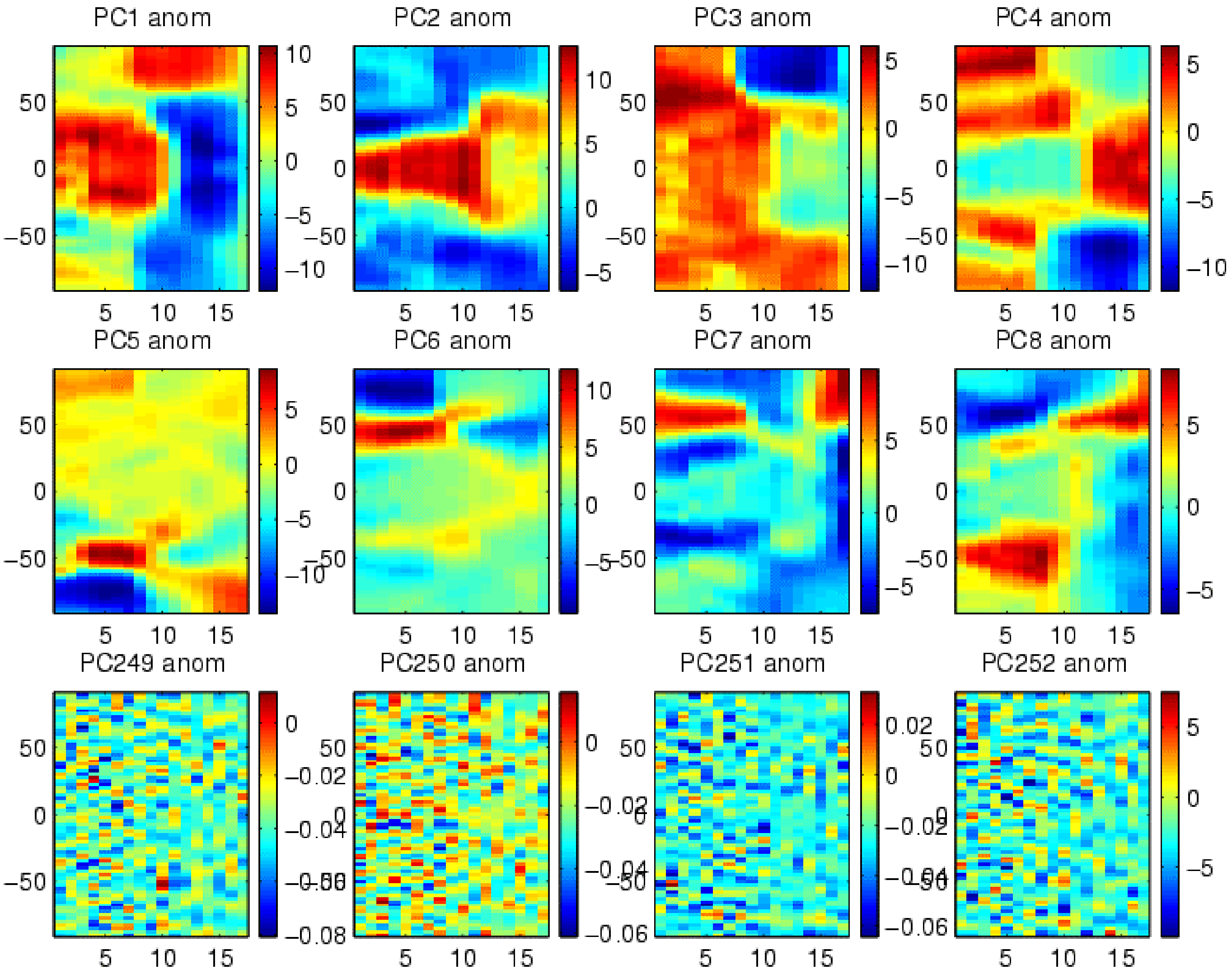
CASC

# Example ICs for the zonal anomaly data



**Not clear how to interpret the estimates. They are independent, but do not correspond to known physical phenomena.**

# Example PCs (cov matrix) for the zonal anomaly data



| #PC | Cumulative %Variation |
|---|---|
| 1 | .15 |
| 5 | .46 |
| 10 | .66 |
| 25 | .88 |
| 50 | .96 |
| 252 | 1 |

**Interpretation much more straightforward. Ben Santer was very pleased when we showed him our results, and suggested further analyses.**

CASC

# Future plans

- What do you expect to achieve by Feb. 2003? What are your goals?
- What are your plans for achieving these goals?
- Why are these goals important? To whom?
- Which scientific domain and who are you working with as your token application scientist?
- Why is your work significant? Who will use it?
- How does your work compare with or differ from similar work by others? Why not simply adopt other people’s work in you domain?
- What is your vision at the end of three years? Do you believe you can achieve that? Why?
- Do you think there will be unsolved problems in your domain at the end of three years? What would you plan to propose?

# What do you expect to achieve by Feb. 2003? What are your goals?

- Follow up on our discussion with Ben Santer
  - look at the PCA time series for covariance and correlation matrices for zonal means
  - incorporate post-processing suggested by Ben
  - correlate with ENSO, volcano signals, and other time series
  - investigate alternative ICA implementations
  - summarize in a report
- See if possible to incorporate constraints in the ICA to separate non-linearly mixed signals: risky
- Complete a design of the ICA implementation in C++, incorporating our enhancements

➔ Goal: help to improve the climate scientist’s understanding of how the signals can be separated

# What do you expect to achieve by Feb. 2003? What are your goals? (contd.)

- Several aspects of our scientific discovery work are high risk
  - poor understanding among climate scientists on how the various signals interact
  - not always easy to interpret the output from ICA
  - existing techniques not always well understood
  - techniques work in some cases but not in others
- We may not be able to solve the entire problem
  - but, any progress is valued by Ben Santer
  - even a negative result!
  - still an important problem that generates great interest
- The techniques are very specific to this problem

➔ Scientific discovery is hard!

# What are your plans for achieving these goals?

- Understand and implement the post-processing needed for the PCs
- Convert the PCA “images” into time series
- Implement the correlation between the PCs and the various signals to see if we can determine which PC represents what
- Investigate new ICA implementations that give more “meaningful” ICs
- ICA with constraints (risky)
  - literature search
  - software implementation in Matlab

# Why are these goals important? To whom?

- They help us to better understand the behavior of the earth's temperature when naturally occurring phenomena are removed
  - identify the contributions of man-made sources
  - understand global warming
  - make better comparisons of climate models
- A better understanding of how the signals interact and can be removed is of interest to climate scientists such as Ben Santer

# Which scientific domain, and who is your application scientist?

- Climate
- Ben Santer, Program for Climate Model Diagnosis and Intercomparison (PCMDI)
  - MacArthur fellowship for research supporting the finding that human activity contributes to global warming
- Future (beyond this domain): HEP, working with LBNL

# Why is your work significant? Who will use it?

- Our work helps in better understanding of the separation of sources contributing to the temperature
- For our work so far, we expect that our findings will contribute to a better understanding of climate models and global warming
- The results will be used by climate scientists
- Future (beyond Feb'03):
  - investigate other dimension reduction techniques for this problem
  - use the dimension reduction techniques in conjunction with sampling to improve indexing and clustering in HEP data

# How does your work compare with others? Why not simply adopt their work?

- To the best of our knowledge, no one else is looking at techniques such as ICA for the separation of mixed signals in climate data
- The existing techniques for this problem are simplistic and  involve knowing something about the kinds of signals that are mixed
- Our approach tries to find the signals in the mix without knowing what kinds of signals they are
- Future (beyond this problem)
  - No one else is looking at effective sampling to improve the efficiency of dimension reduction

# What is your vision at the end of three years? Do you believe you can achieve it?

- Scientific discovery
  - —Investigate more complex mixing models
  - —understand how much PCA, ICA, and related techniques can contribute to the separation of the signals
  - —issue: to determine when we have reached the point of diminishing returns
- Software tools
  - —for PCA, ICA, and related techniques
  - —with sophisticated sampling for large data sets
- A better understanding of how these techniques apply to real datasets in climate and high energy physics

# Will there be unsolved problems at the end of 3 years? What will you propose?

- Definitely!
- Climate, high energy physics and other applications are replete with data analysis problems
  - application of dimension reduction techniques
  - analysis of time series data
  - analysis of HEP data