6 April 2007

Crossing Disciplinary Boundaries: Novel Techniques for Data Analysis in Space Physics

Dmitri Kondrashov

University of California, Los Angeles

Motivation

1. Geophysical time series have typically broad peaks on top of a continuous, "warm-colored" background ➔ ***Method***

2. Connections to dynamics ➔ ***Theory***

3. Need for stringent statistical significance tests ➔ ***Toolkit***

4. Applications to analysis and prediction ➔ ***Examples***

Joint work with M. Ghil and many others

http://www.atmos.ucla.edu/tcd

Motivation & Outline

1. Data sets in the geosciences are often short, contain noise (errors) and are gappy: this is both an obstacle and an incentive.

2. Phenomena in the geosciences often have both regular ("cycles") and irregular ("noise") aspects.

3. Different spatial and temporal scales:
   one person's noise is another person's signal.

4. Need both deterministic and stochastic modeling.

5. Regularities include (quasi-)periodicity ➔ spectral analysis via "classical" and novel methods – singular spectrum analysis (SSA).

6. Reconstruction of gappy data with SSA.

7. Does some combination of the two, + deterministic and stochastic modeling, provide a pathway to prediction?
   Empirical model reduction

8. Be prepared to answer questions...

For details and publications, please visit:

TCD http://www.atmos.ucla.edu/tcd/

Spatio-Temporal Variability

- Standard view — Binary thinking:

  Trend — Predictable (completely), deterministic, reassuring, good;

  Variability — Unpredictable (totally), stochastic, disconcerting, bad.

- In fact, these two are but extremes of a spectrum of, more or less predictable, types of behavior, between the totally boring & the utterly surprising.

- (Linear) Trend = Stationary >

  Periodic > Quasi-periodic >

  Deterministically aperiodic >

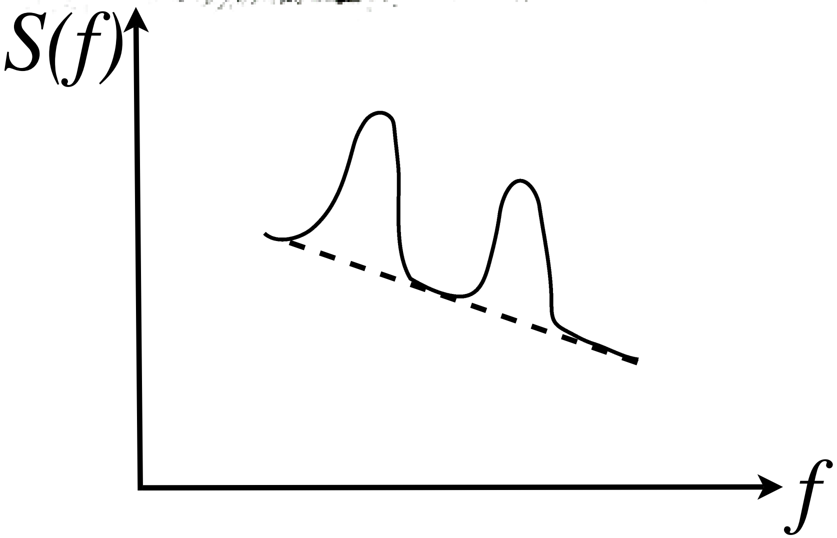
  Random Noise

- Here ">" means "better, more predictable", &

  Variability = Trend+ Periodic + Quasi-periodic +

  Aperiodic + Random

Spectral Density (Math)/Power Spectrum (Science & Engng.)



Variance vs. frequency

Continuous background
+ peaks (poles)

- **Wiener-Khinchin Theorem <-> Blackman-Tukey Correlogram**

$$R(s) = \lim_{L\to\infty} \frac{1}{2L} \int_{-L}^{L} x(t)x(t+s)dt$$

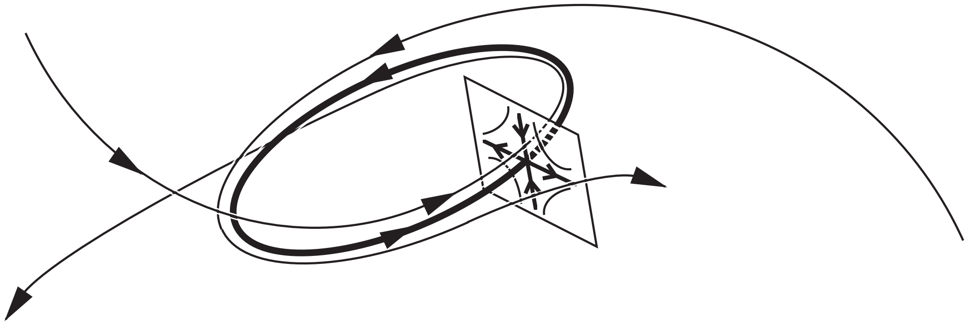
$$S(f) = \frac{1}{2\pi} \int_{-\infty}^{\infty} R(s)e^{-ifs}ds \equiv \hat{R}(s)$$

Time-domain<->frequency domain: **lag-autocorrelation function** & the **spectral density** are **Fourier transforms** of each other.

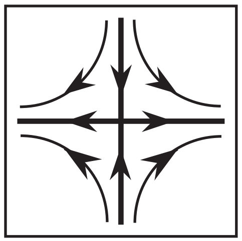
Power Law for Spectrum (cont'd)

- Nonlinear climate hypothesis: "***Poles***" correspond to the least unstable periodic orbits

"*unstable limit cycles*"



"*Poincaré section*"



- Major clue to the physics that underlies the dynamics
- Orbits are not necessarily elliptic, i.e. not

$$(x, y) = (a_f sin(ft), b_f cos(ft))$$

- but phase and amplitude modulation and intermittent behavior.

$$(x, y) = (a_f(t) sin(ft + \phi(t)), b_f(t) cos(ft) + \Psi(t))$$

Power Law for Spectrum

$$S(f) \sim f^{-p} + \textit{poles}$$

i.e. linear in log-log coordinates

For a 1st-order Markov process or **"red noise" $p = 2$**

"Pink" noise, $p = 1$ ($1/f$, flicker noise)

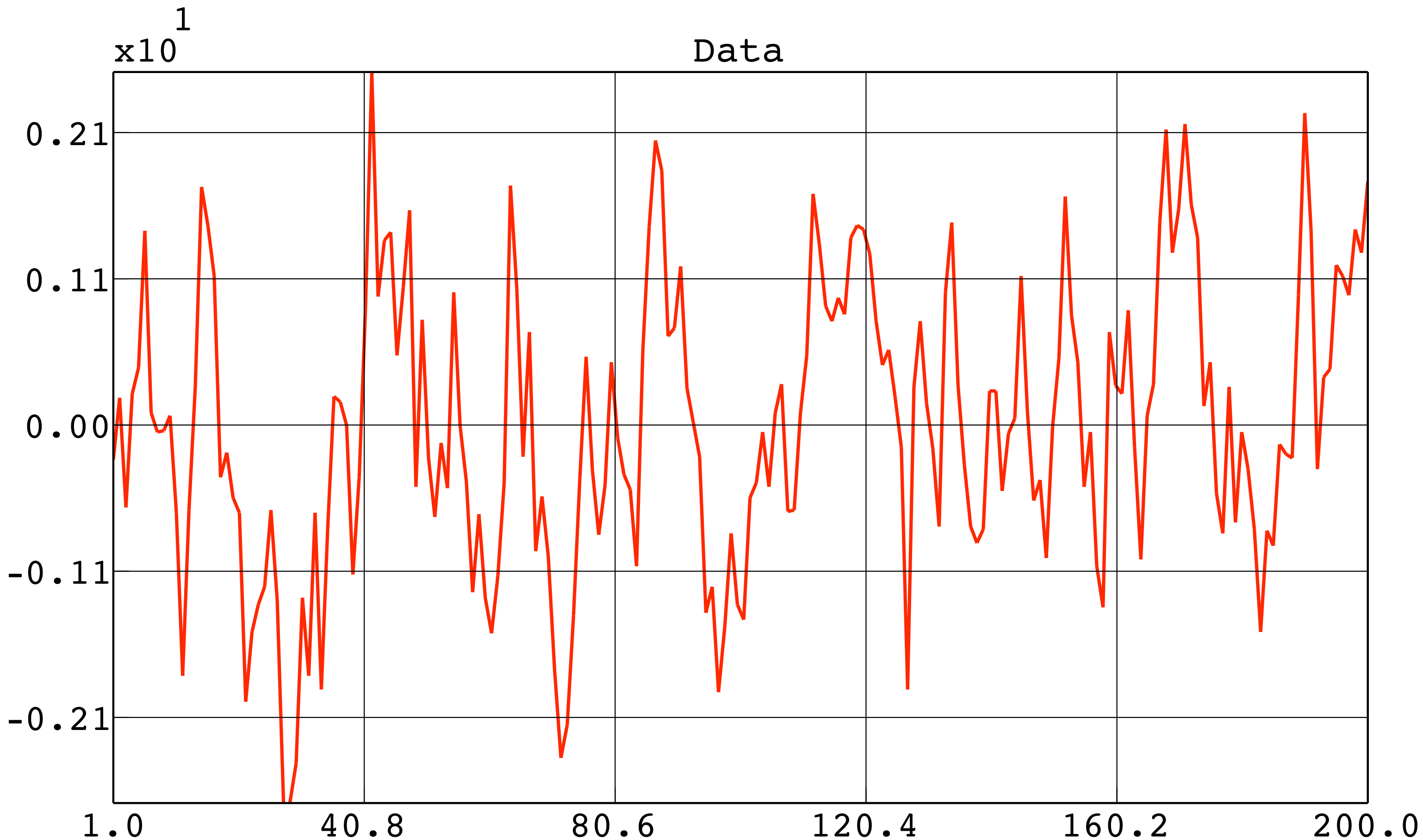
"White" noise, $p = 0$

It is a challenge for ***short and noisy*** geophysical time series to distinguish between **poles** and **red noise.**

$$\ddot{x} = -\omega^2 x \;\; vs. \;\; \dot{x} = -\lambda x$$

Tradeoff for spectral methods: **resolution (spurious peaks) vs. robustness (power leakage)**

Synthetic example

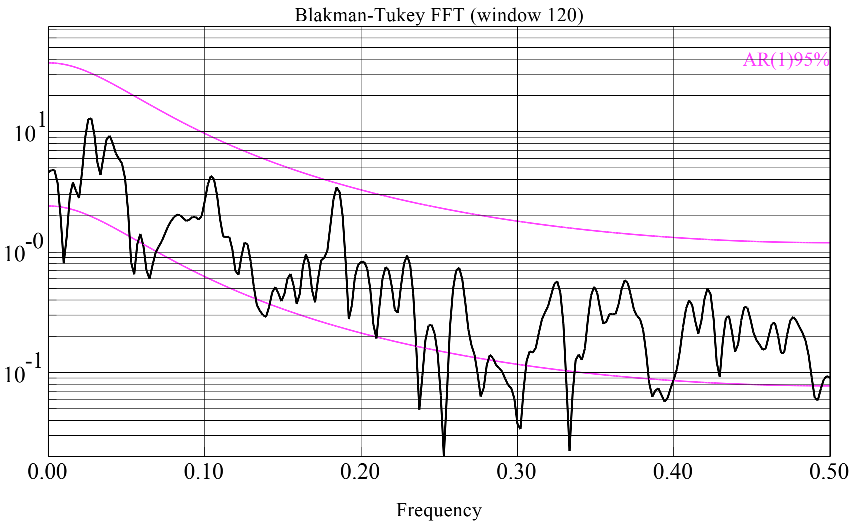


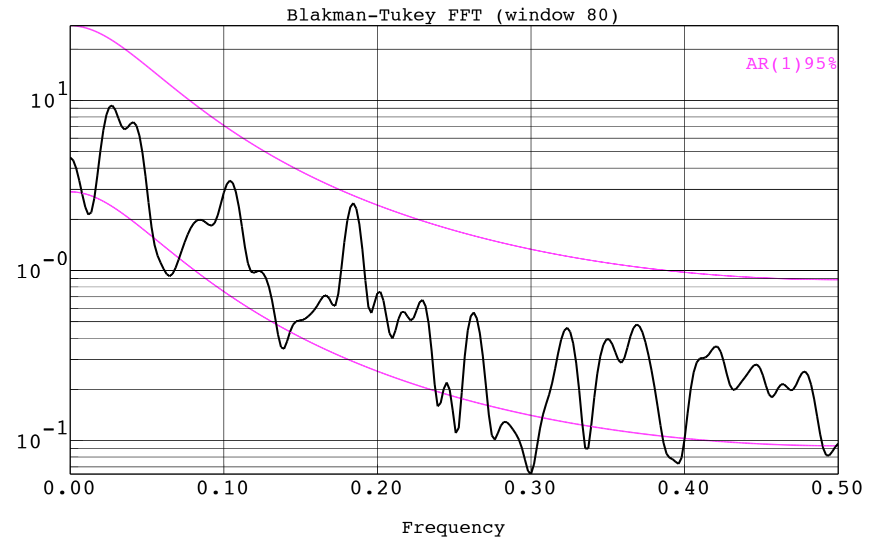
Q: ***Is there a periodicity and what is its frequency?***

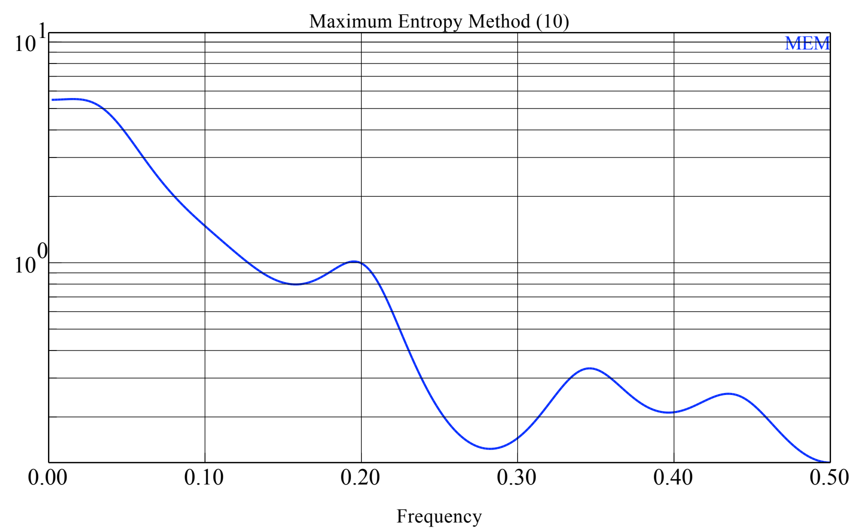
Hint: ***It is a periodic signal contaminated by noise...***

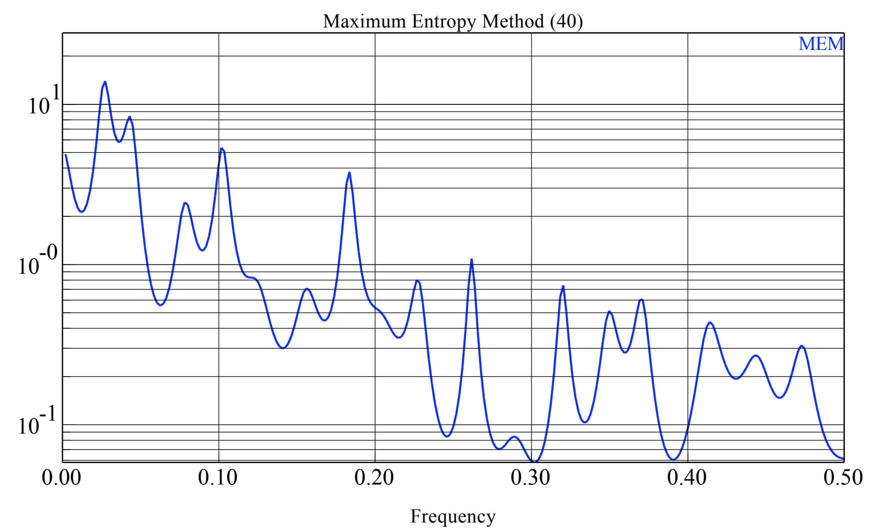
A: ***What is the underlying noise "null hypothesis"?***

Classical Spectral Methods









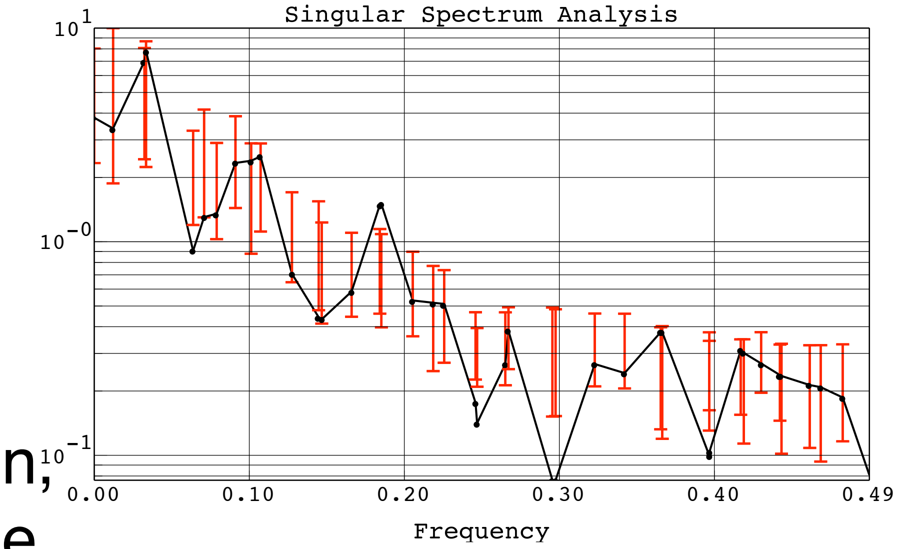
Advanced Spectral Methods

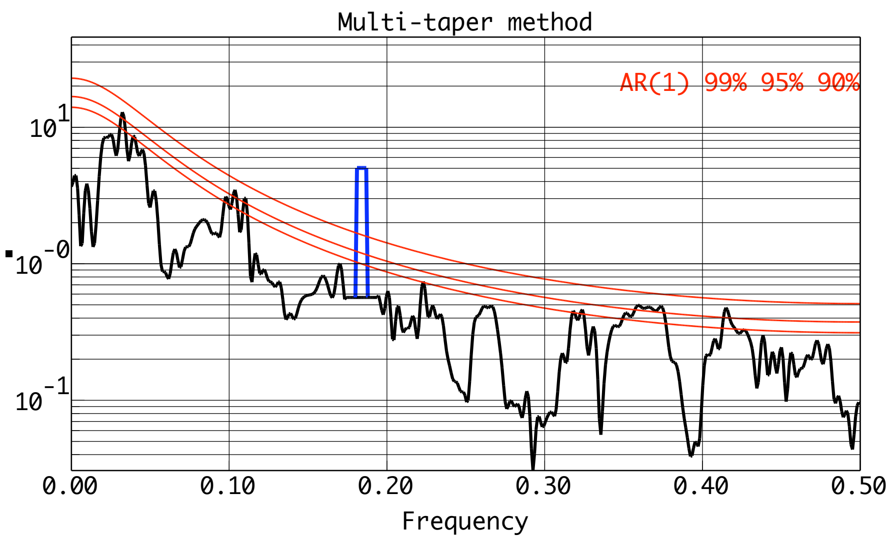
- **Singular spectrum analysis (SSA)** and **Multi-taper method (MTM)**.

- detection of periodic signals: phase and amplitude modulation, intermittent behavior, large noise.

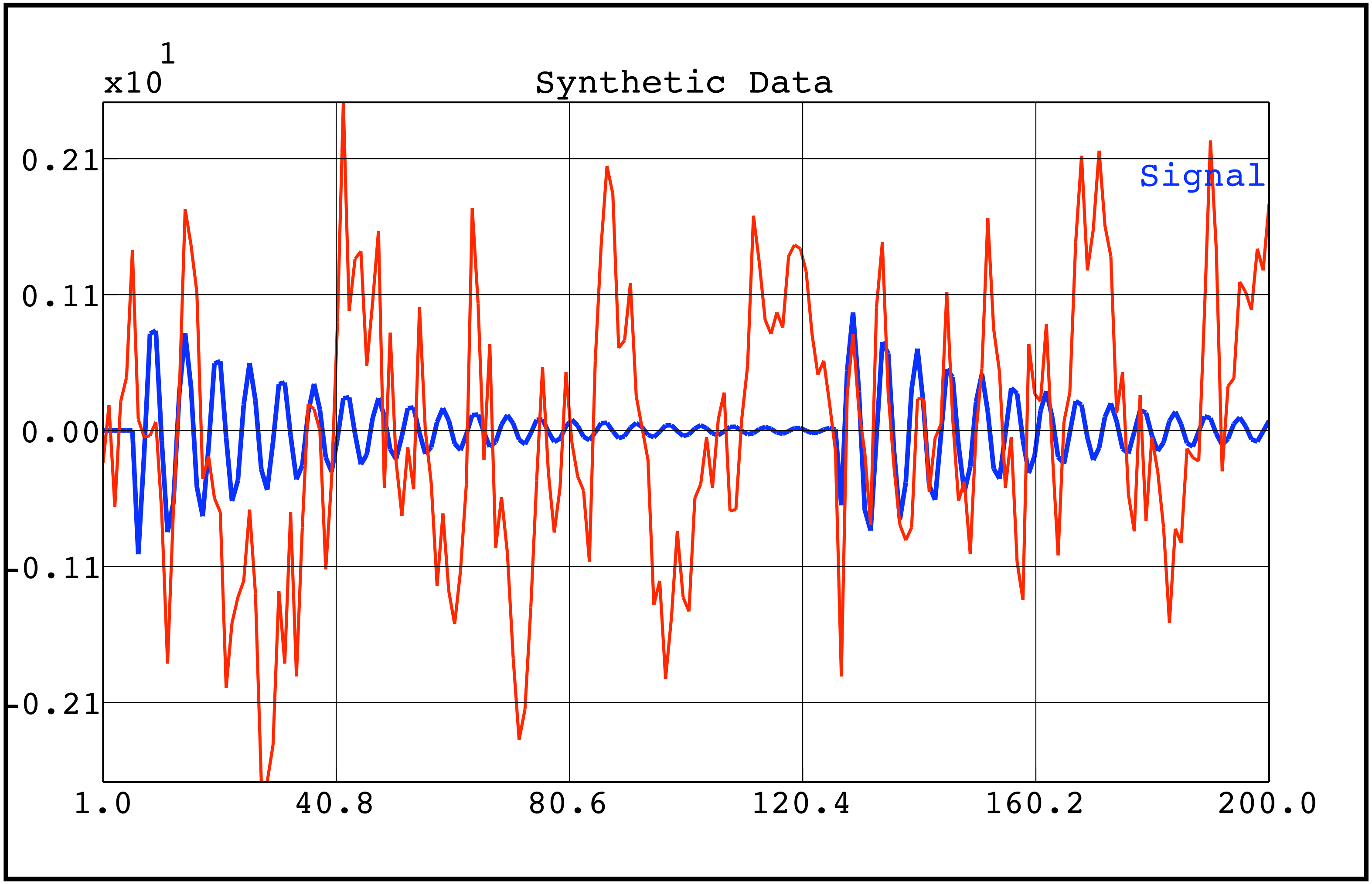
- use **data-adaptive** orthogonal basis in **frequency domain** (MTM) and **time domain** (SSA).

- significance tests for spectral peaks.





Anybody guessed it right?



Singular Spectrum Analysis (SSA)

Spatial EOFs, Principal Component Analysis (PCA)

x -- space

$$\phi(x,t) = \sum a_k(t)e_k(x)$$

$$C_\phi(x,y) = E\phi(x,\omega)\phi(y,\omega) = \frac{1}{T}\int_o^T \phi(x,t)\phi(y,t)\,dt$$

$$C_\phi e_k(x) = \lambda_k e_k(x)$$

Spatio-temporal EOFs, SSA

s -- lag

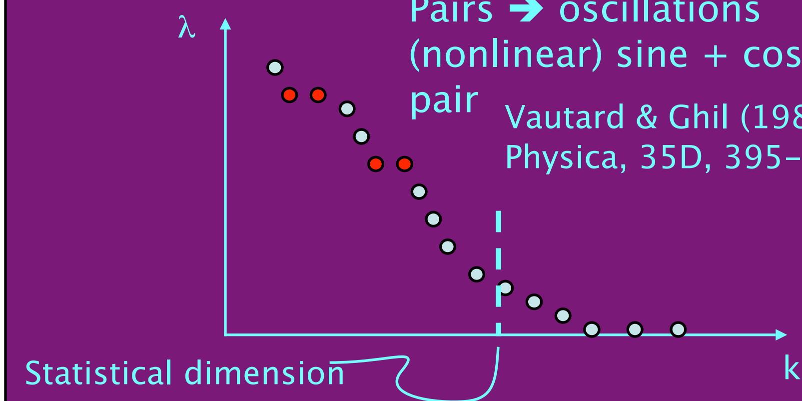
$$X(x+s) = \sum a_k(t)e_k(s)$$

$$C_X(s) = EX(t+s,\omega)\phi(s,\omega) = \frac{1}{T}\int_o^T X(t)X(t+s)\,dt$$

$$C_X e_k(s) = \lambda_k e_k(s)$$

Pairs → oscillations (nonlinear) sine + cosine pair

Vautard & Ghil (1989: VG) Physica, 35D, 395–424



Empirical Orthogonal Functions (EOFs) are the most **optimal patterns** to **capture the variance**.

EOFs are **statistical** features, but may describe some **dynamical** (physical) mode(s) in low-order dynamical systems

SSA Power Spectra & Reconstruction

- **A. Transform pair**:

$$X(t+s) = \sum_{k=1}^{M} a_k(t) e_k(s),\, e_k(s) - EOF$$

For given window ***M***, e_k's are **adaptive filters** (empirical orthogonal functions)

$$a_k(t) = \sum_{s=1}^{M} X(t+s) e_k(s),\, a_k(t) - PC$$

the a_k's are **filtered time series**, principal components in time domain.

B. Power spectra

$$S_X(f) = \sum_{k=1}^{M} S_k(f);\quad S_k(f) = \hat{R}_k(s);\quad R_k(s) \approx \frac{1}{T}\int_0^T a_k(t) a_k(t+s)\,dt$$

C. Reconstruction

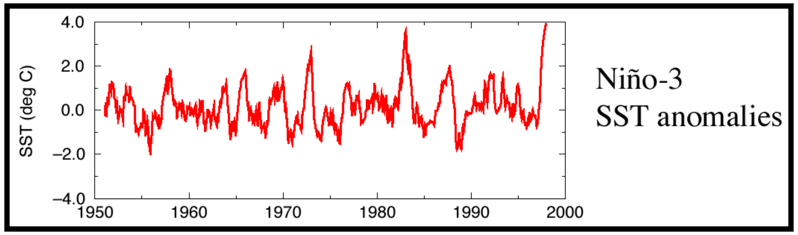
$$X^K(t) = \frac{1}{M} \sum_{k \in K} \sum_{s=1}^{M} a_k(t-s) e_k(s);$$

in particular: $K = \{1, 2, \ldots.., S\}$ or $K = \{k\}$ or $K = \{l, l+1; \lambda_l \approx \lambda_{l+1}\}$

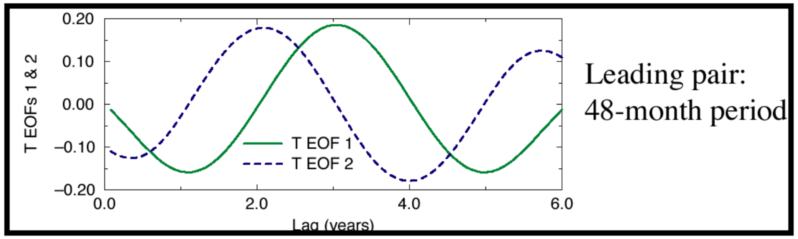
SSA of Nino-3 index (El-Nino)

SSA decomposes (geophysical & other)
time series into
Temporal EOFs (T-EOFs) and
Temporal Principal Components (T-PCs),
based on the series' lag-covariance matrix

Time series

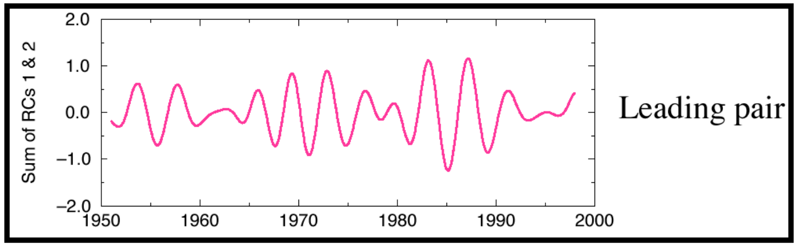


T–EOFs



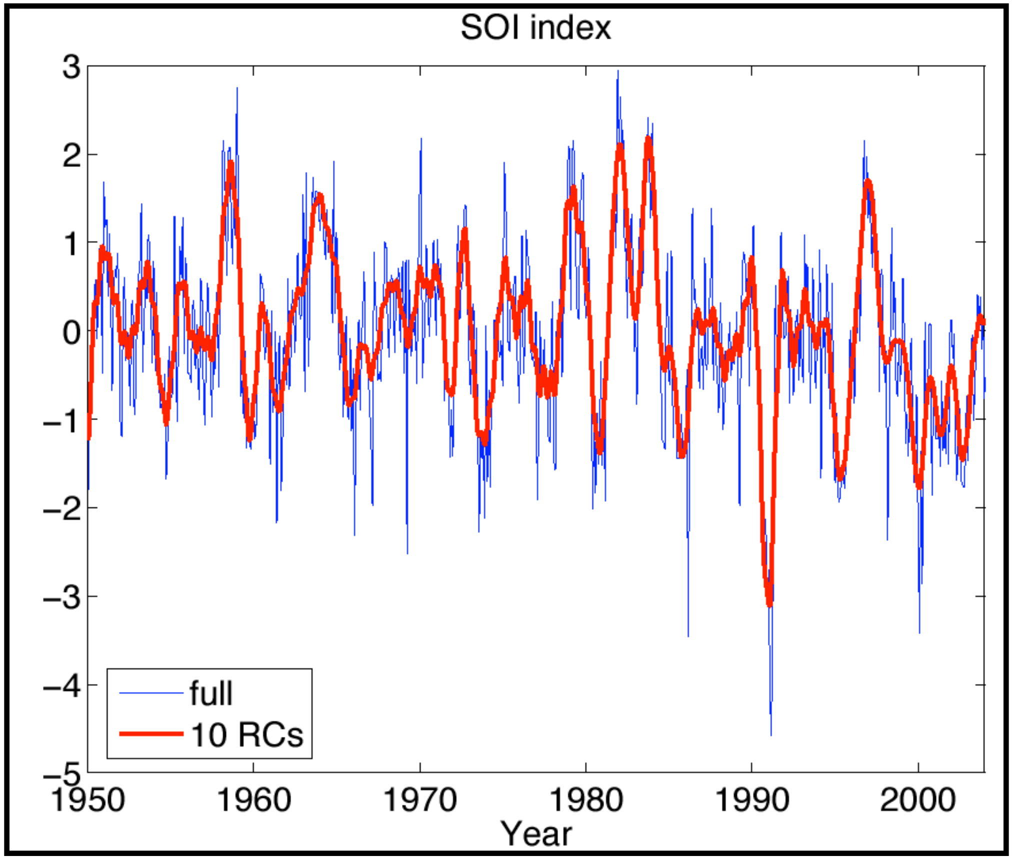
Selected parts of the series can be
reconstructed, via
Reconstructed Components (RCs)

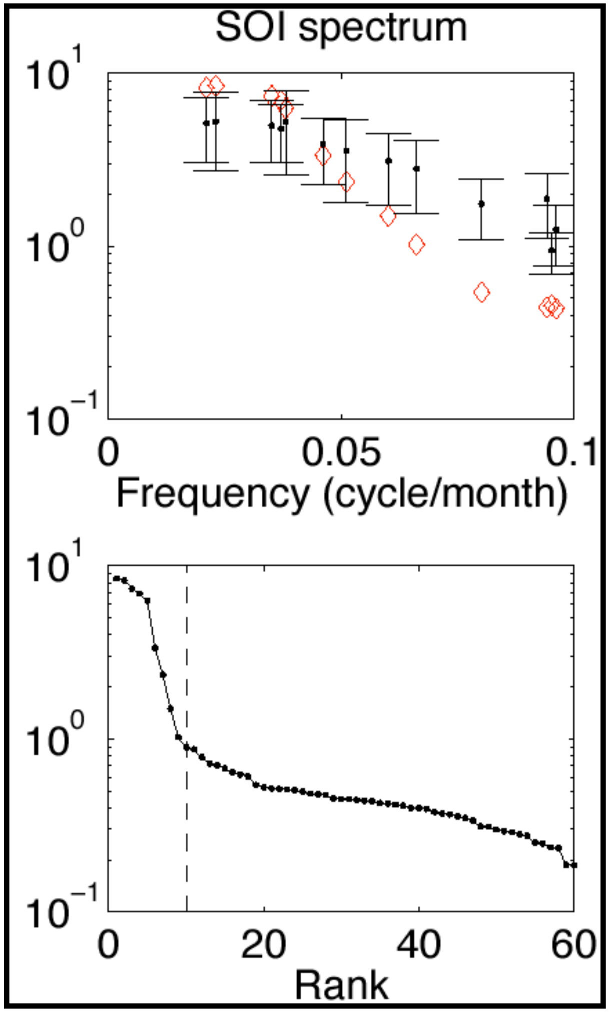
RCs



- SSA is good at isolating oscillatory behavior via paired eigenelements.
- SSA tends to lump signals that are longer-term than the window into
  - –one or two trend components.

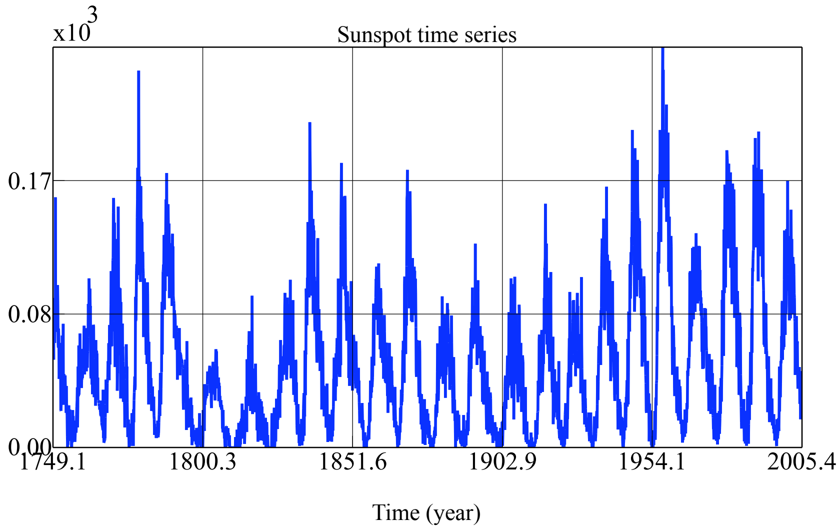
SSA of Southern Oscillation Index (El-Nino)

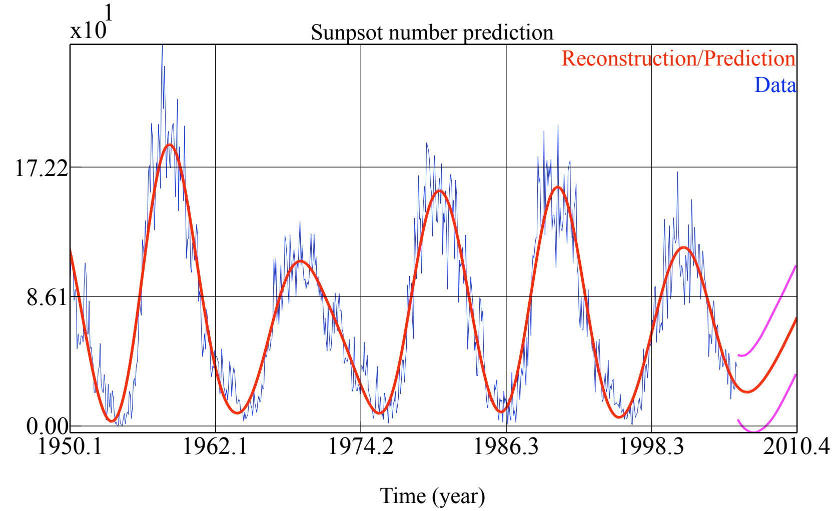




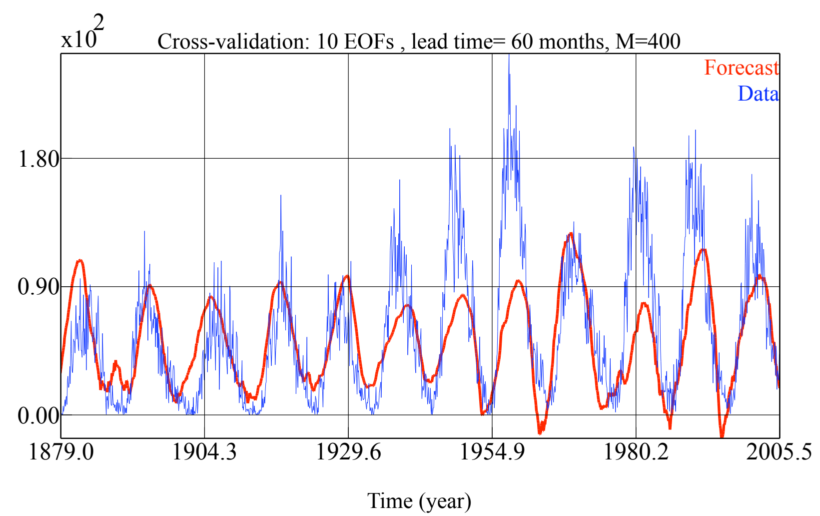
- **Powerful noise filter**: Break in slope of SSA spectrum distinguishes **"significant"** from **"noise"** EOFs
- Formal Monte-Carlo test identifies 4-yr and 2-yr ENSO oscillatory modes (**SSA pairs**). A window size of M = 60 is enough to "resolve" these modes in a monthly SOI time series.

SSA Forecast (Sunspot cycle)

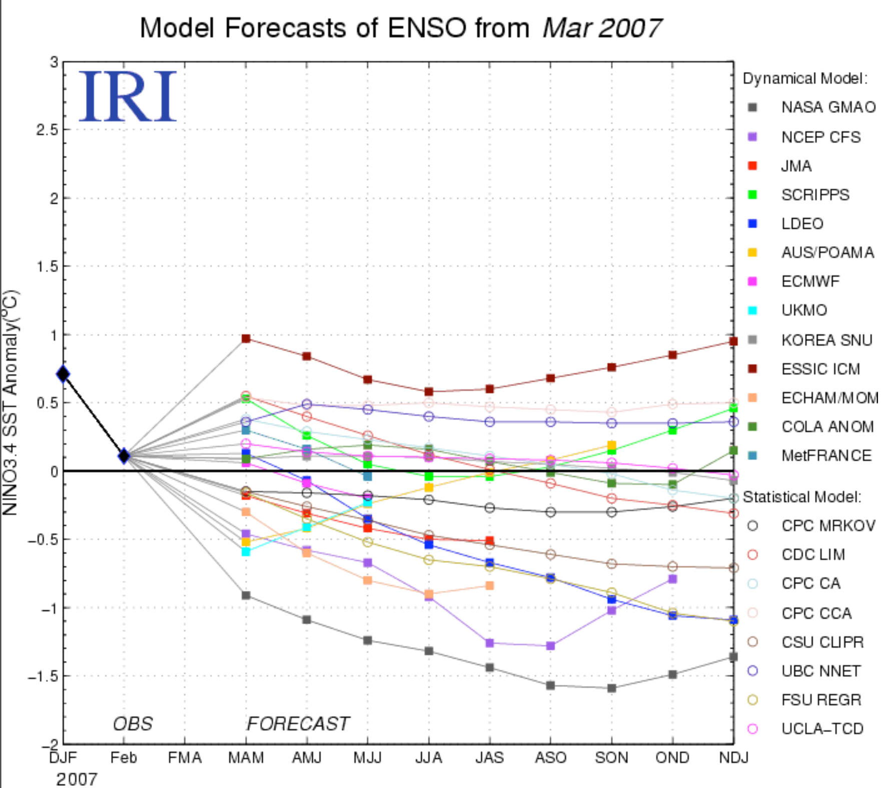


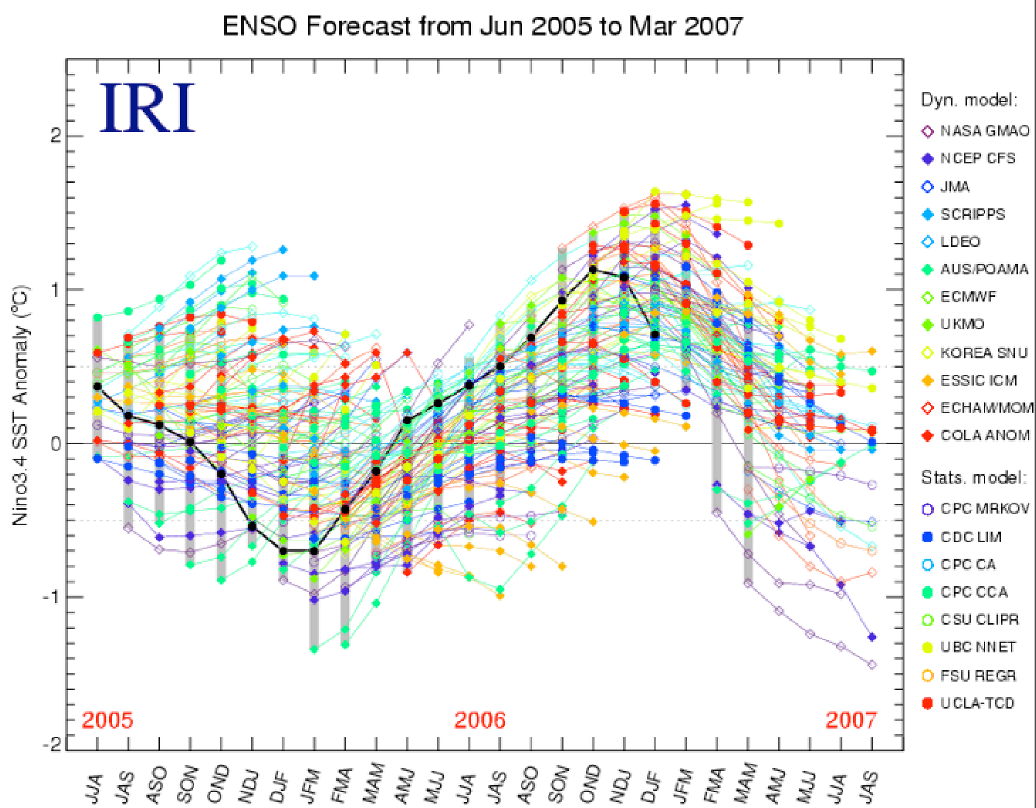


- Forecast principal components of "**signal**" with AR(M) model and do reconstruction.
- Perform cross-validation to find optimum number of "**signal**" components.
- Correlations are both advantage and limitations of empirical models.
- Can be improved with multivariate series.



Future of "Space Weather"?





- Forecast of Nino-3 index 1-yr ahead, and recent performance.
- Real-time forecasting is tough even with many good models and plentiful observations!

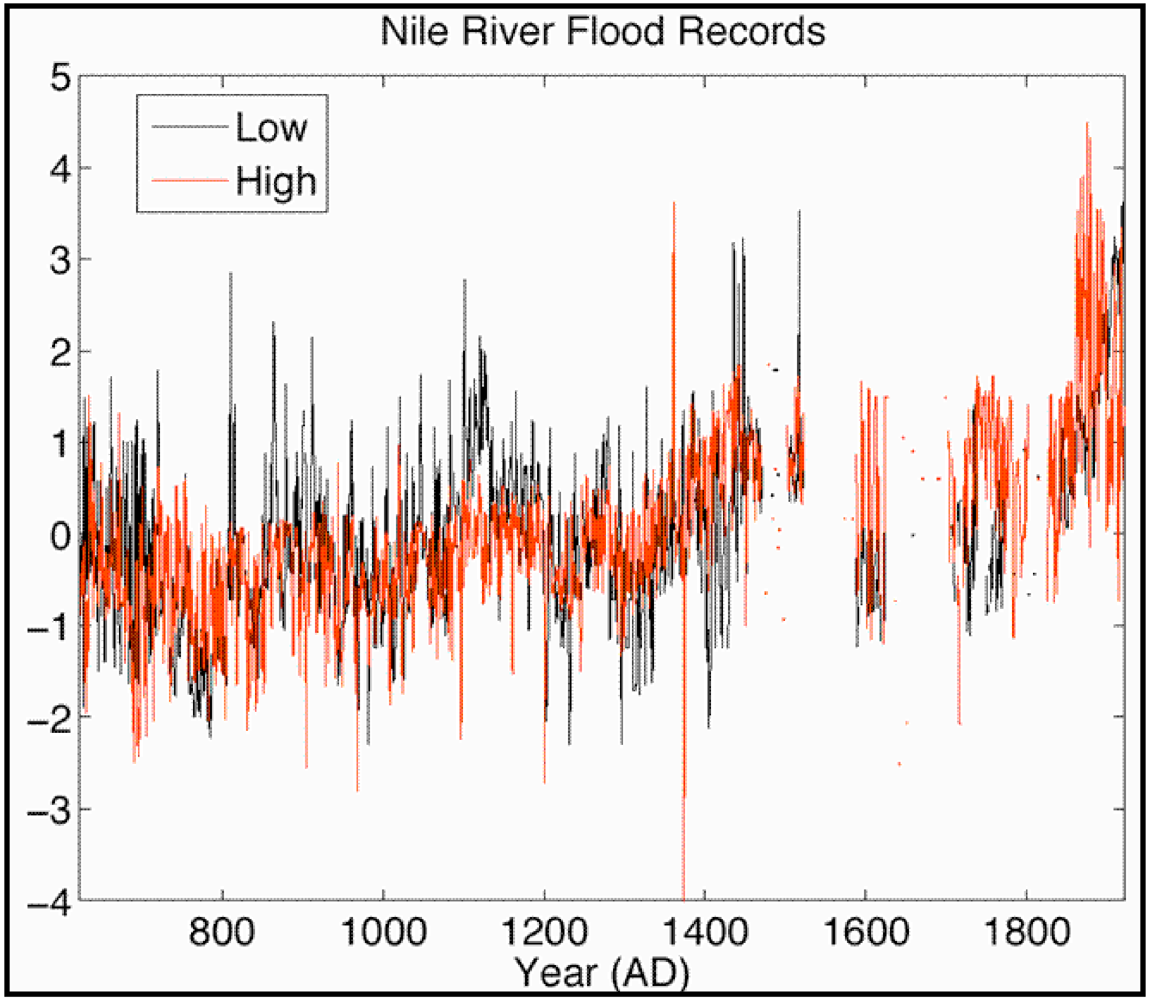
Dealing with

Missing

Data

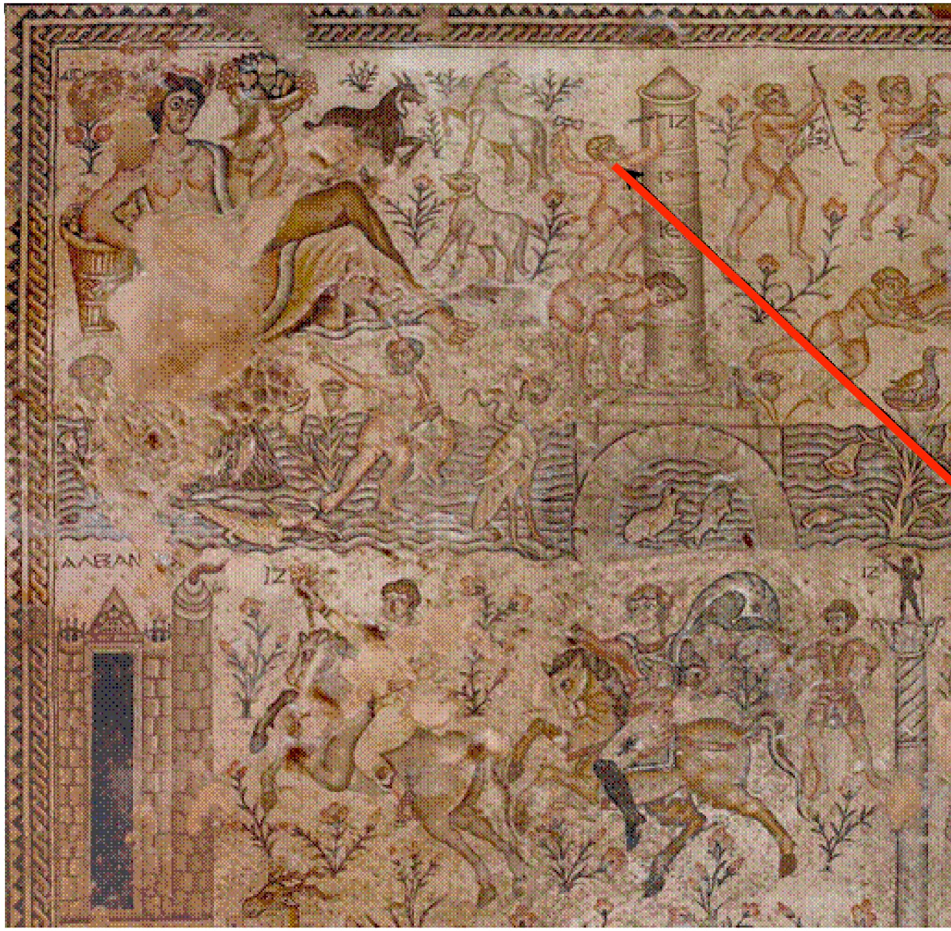
w/o data assimilation

Historical records are full of "gaps"....



Annual maxima and minima of the water level at the nilometer on Rodah Island, Cairo.

Why are there data missing?



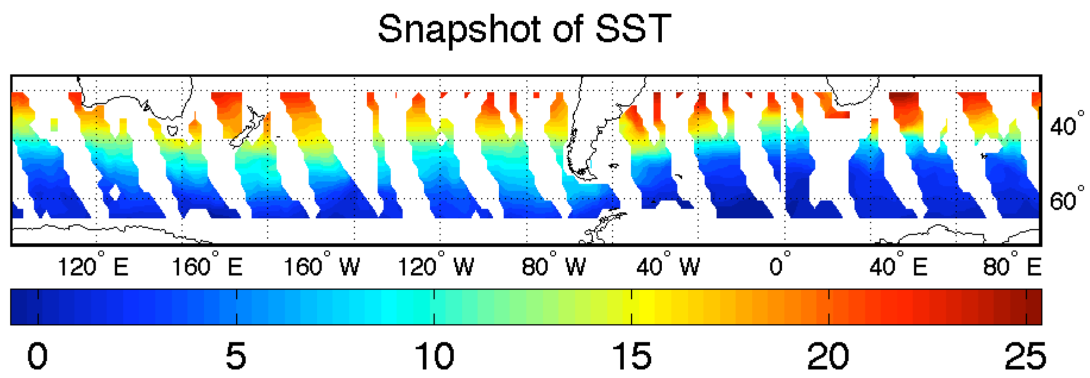
Hard Work

- Byzantine-period mosaic from Zippori, the capital of Galilee (1st century B.C. to 4th century A.D.); photo by Yigal Feliks, with permission from the Israel Nature and Parks Protection Authority)
- Is there 14-yr cycle there (fat and lean years?)

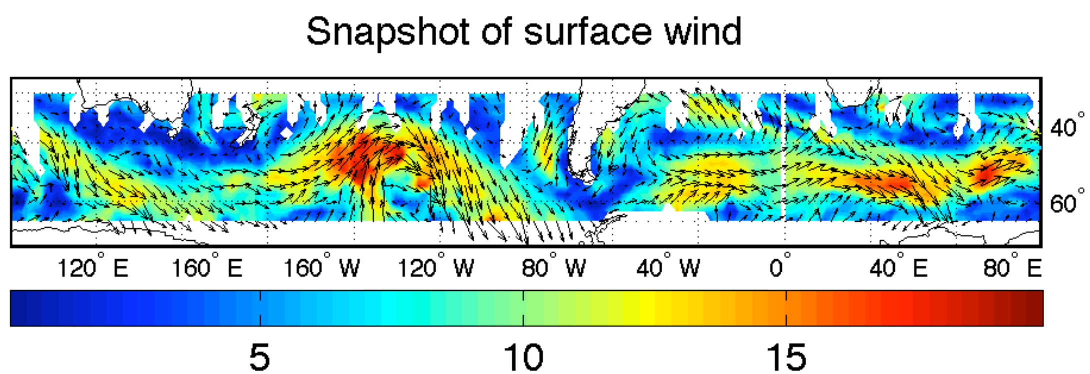
... and now on Earth...

- SST (AMSR-E), daily 2x2, June 2002 – February 2007: 38.2% of missing points
- Wind (QuikSCAT), daily 2x2, July 1999 -- February 2007:17.2% of missing points

(a)

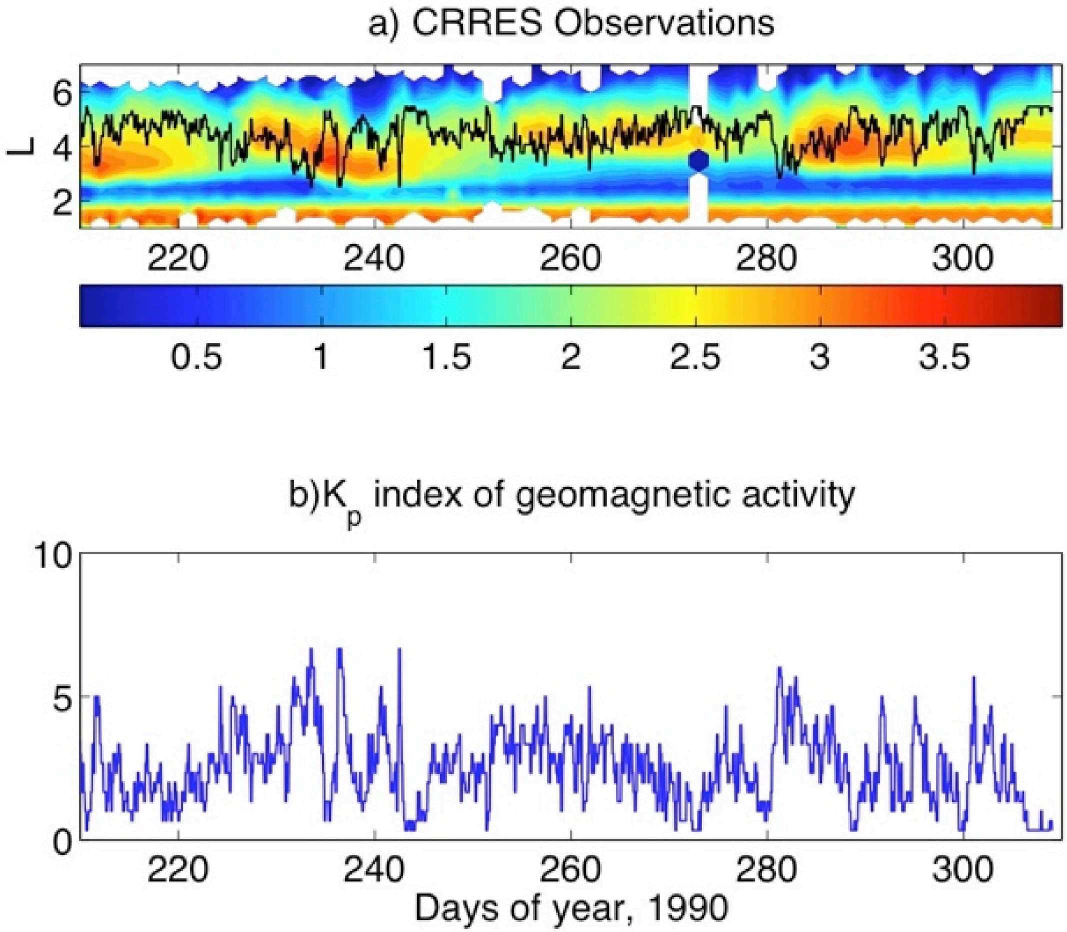


(b)



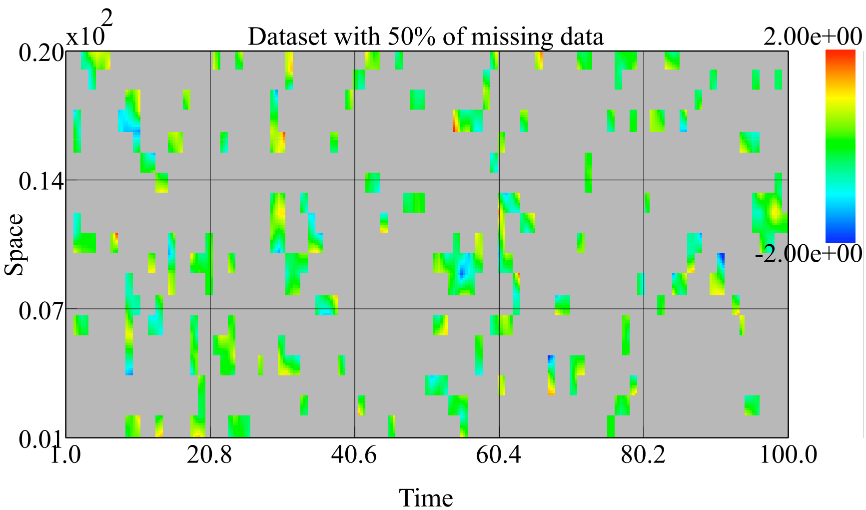
- Gaps: satellite coverage, precipitation and clouds.

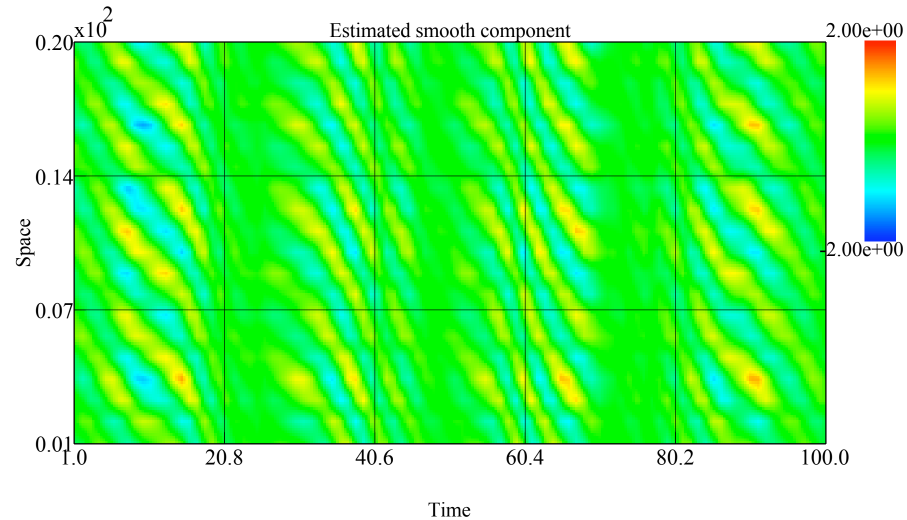
... and in Space!

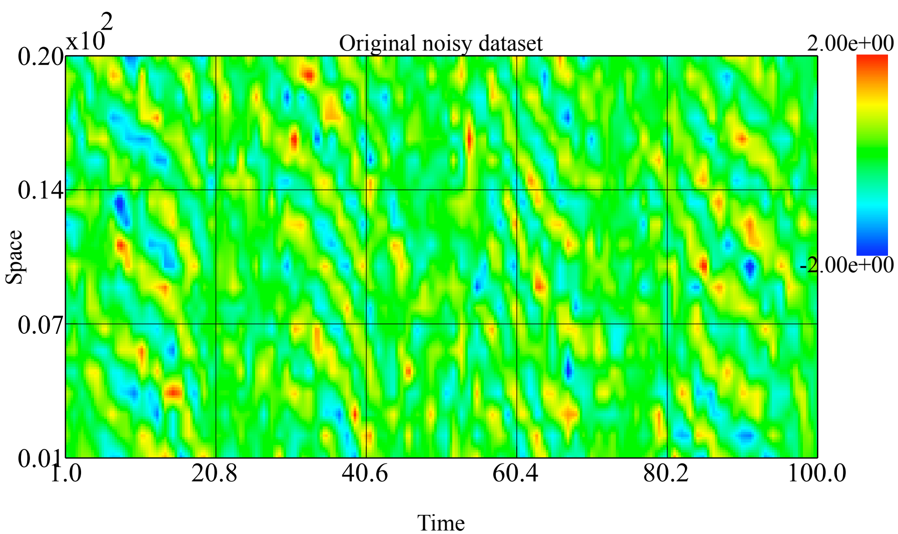


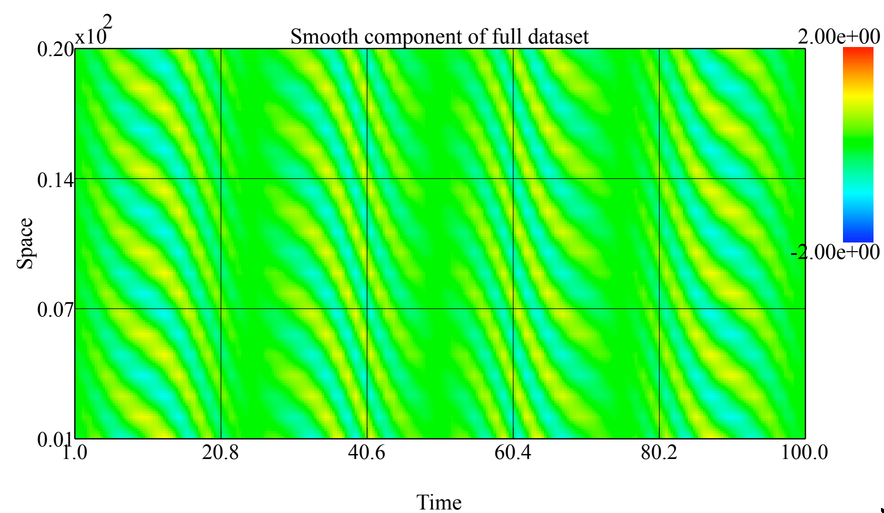
- Gaps: satellite coverage, malfunctions.

How SSA can help with the gaps: synthetic example









SSA gap-filling

1. Choose window ***M*** and set ***K***=1. Flag fraction of dataset ***X(t)(t=1:N)*** as "missing" for cross-validation.

2. Update mean and covariance, find leading ***K*** EOFs

$$\mathbf{D} = \begin{pmatrix} X(1) & X(2) & . & . & X(M) \\ X(2) & X(3) & . & . & X(M+1) \\ . & . & . & . & . \\ X(N'-1) & . & . & . & X(N-1) \\ X(N') & X(N'+1) & . & . & X(N) \end{pmatrix}$$

$$\mathbf{C}_X = \frac{1}{N'}\mathbf{D}^{\mathrm{t}}\mathbf{D}; \mathbf{C}_X E_k = \lambda_k E_k$$

3. Reconstruct missing points using ***K*** EOFs

$$A_k(t) = \sum_{j=1}^{M} X(t+j-1)E_k(j)$$

$$R_{\mathcal{K}}(t) = \frac{1}{M_t}\sum_{k\in\mathcal{K}}\sum_{j=L_t}^{U_t} A_k(t-j+1)E_k(j);$$

4. If convergence, ***K = K +1.*** Check cross-validation error, and Go to Step 2 if necessary.

Utilize both spatial and temporal correlations to iteratively compute self-consistent lag-covariance matrix => can be applied to very gappy data.

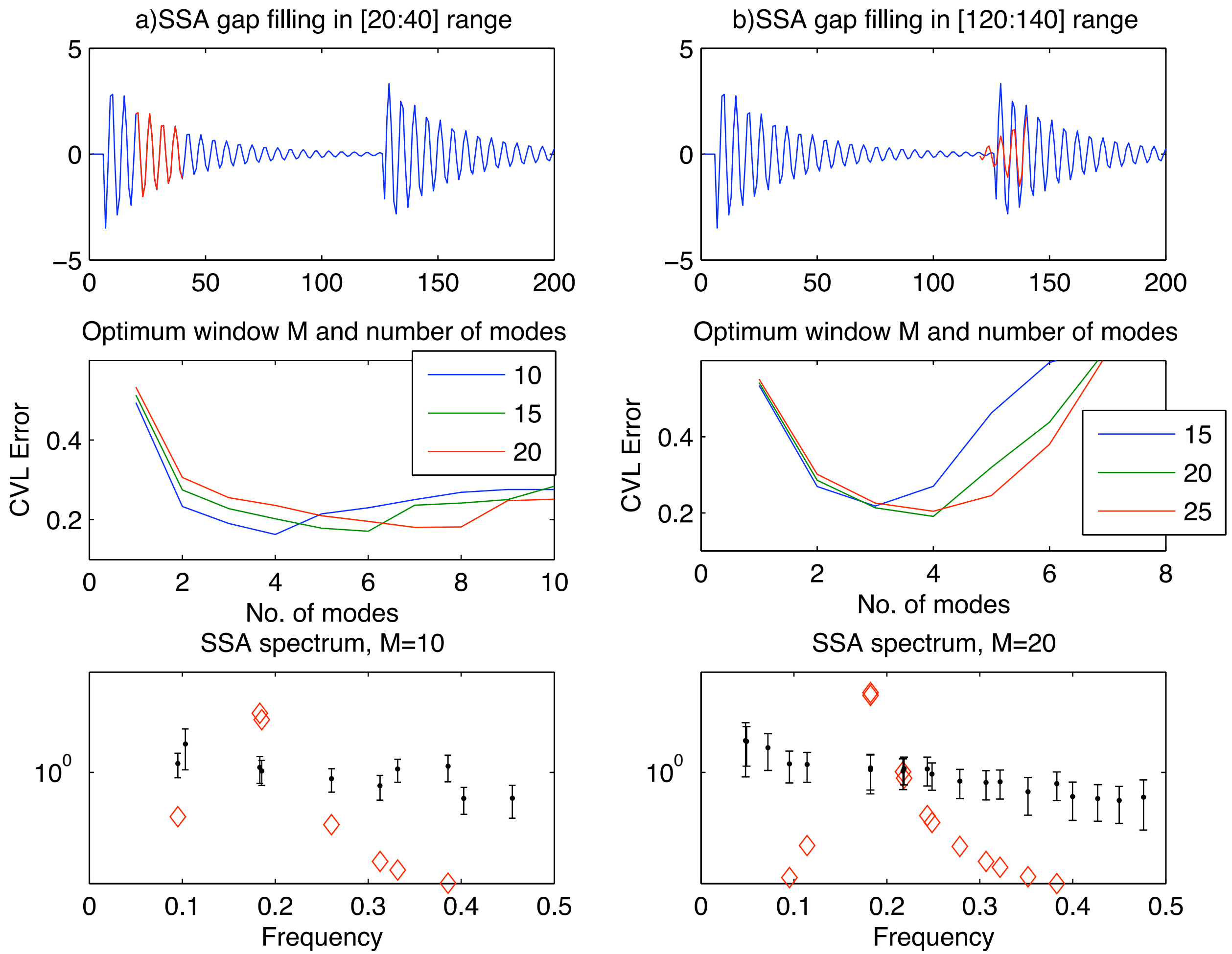
Follows expectation maximization (EM) procedure for finding maximum likelihood estimates of mean and covariance matrix.

A few ***K*** leading EOFs correspond to the "smooth" modes, while the rest is noise.

Provides both spectral analysis and estimates of missing data.

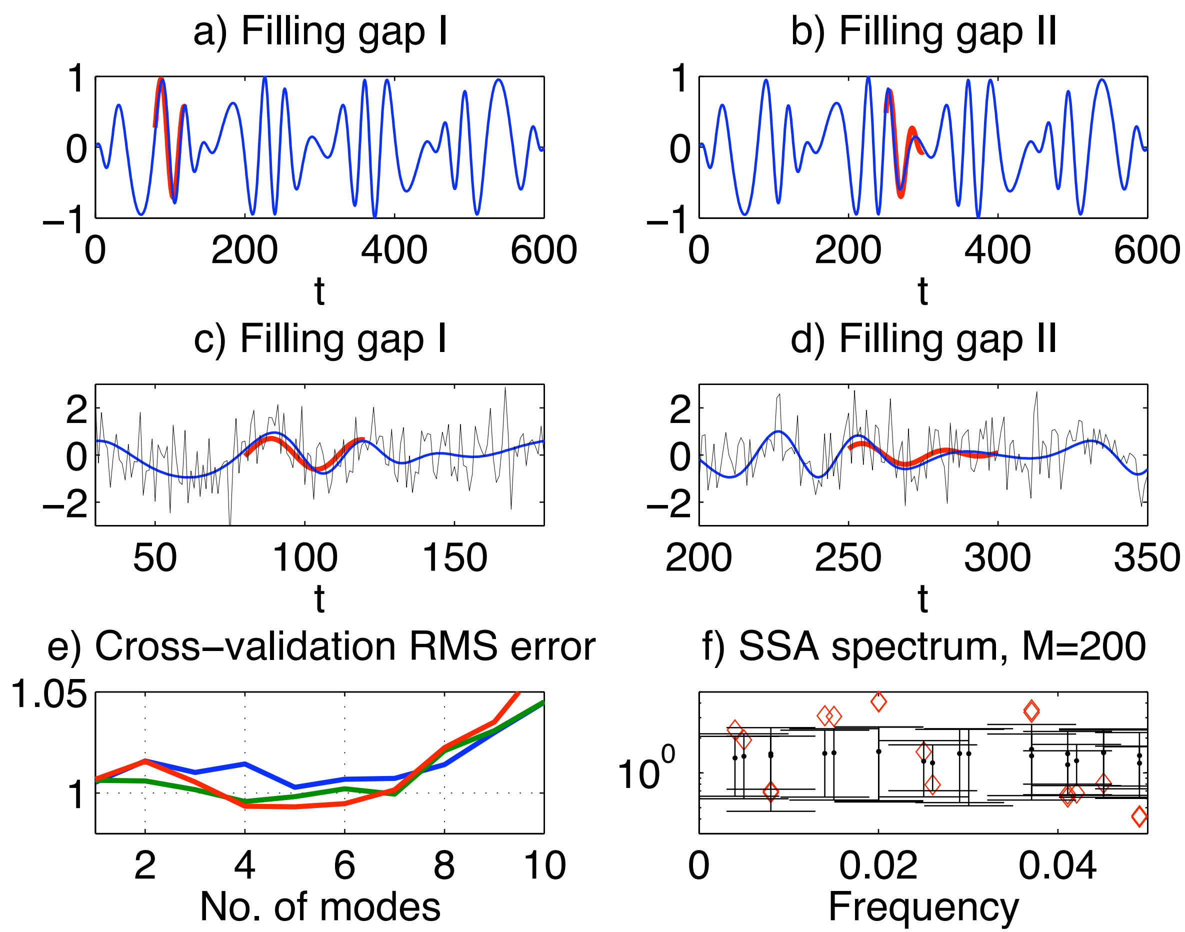
- D. Kondrashov and M. Ghil, 2006: Spatio-temporal filling of missing points in geophysical data sets, Nonl. Proc. Geophys., 13, 151-159.

Synthetic I: Gaps in Oscillatory Signal



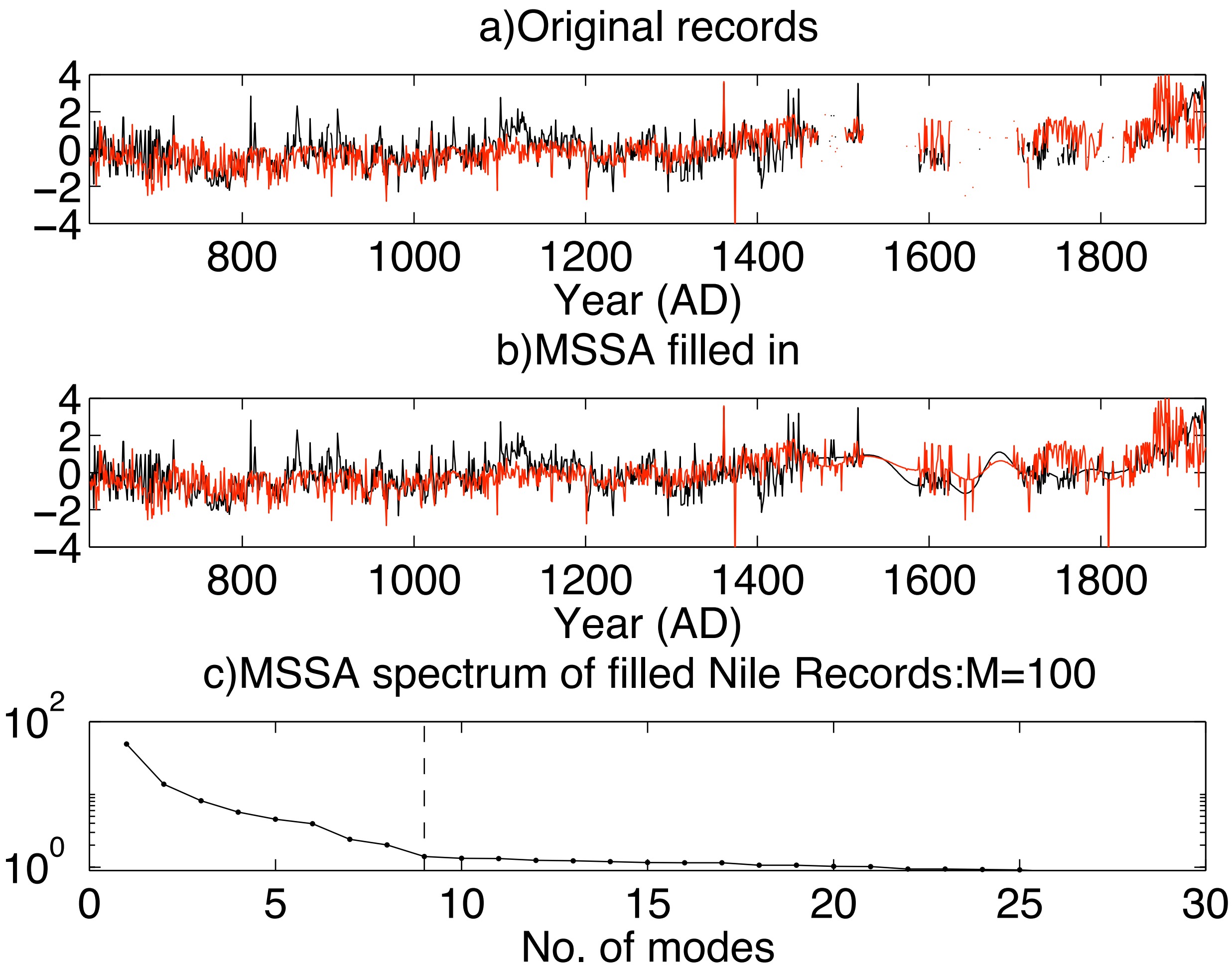
- Very good gap filling for smooth modulation; OK for sudden modulation.

Synthetic II: Gaps in Oscillatory Signal + Noise



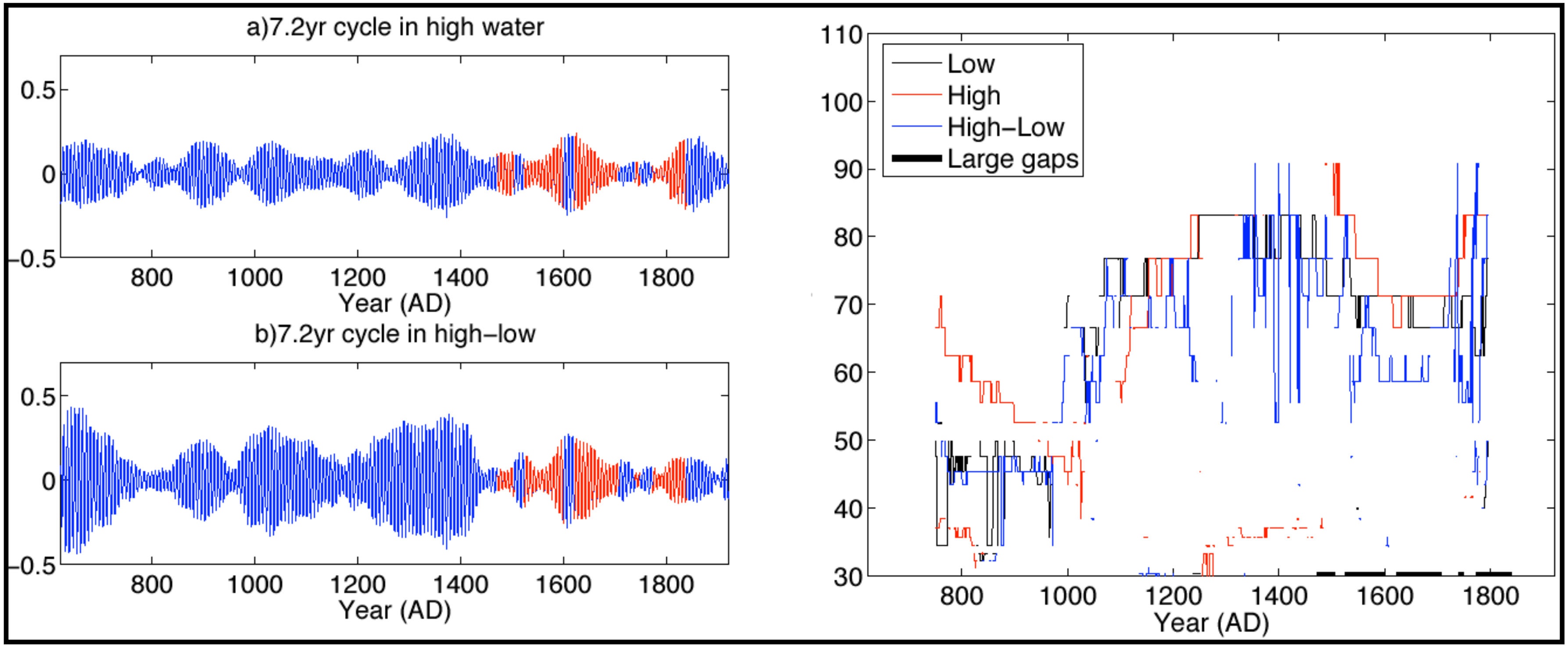
$$x(t) = sin(\frac{2\pi}{300}t) * cos(\frac{2\pi}{40}t + \frac{\pi}{2} sin\frac{2\pi}{120}t)$$

Filed-in Nile River Records



- Kondrashov D., Y. Feliks and M. Ghil (2005): Oscillatory modes of extended Nile River records (A.D. 622-1922), *Geophys. Res. Let.*, 32, L10702, doi:10.1029/2004GL022156

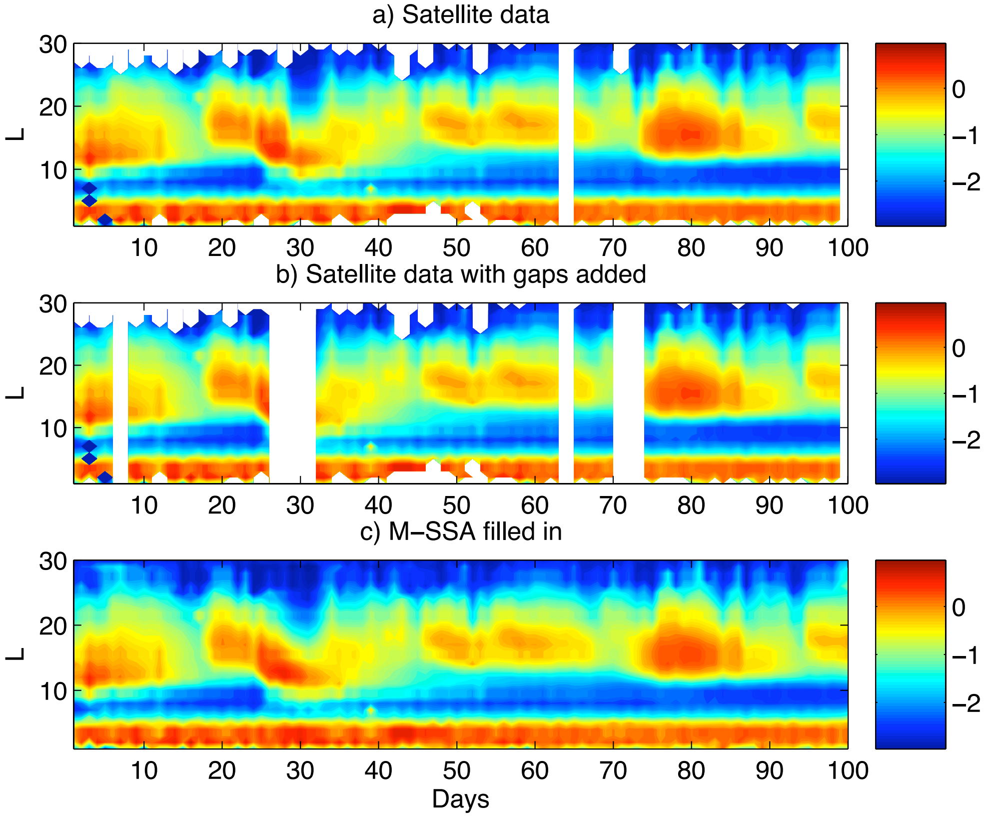
Significant Oscillatory Modes in Nile records

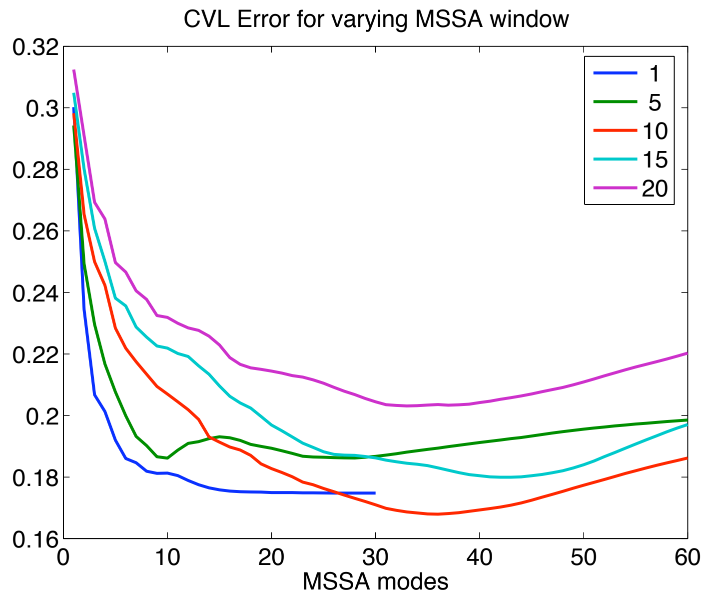


SSA reconstruction of the 7.2-yr mode in the extended Nile River records: (a) high-water, and (b) difference. Normalized amplitude; reconstruction in the large gaps in red.

Instantaneous frequencies of the oscillatory pairs in the low-frequency range (40–100 yr). The plots are based on multi-scale SSA [Yiou *et al*., 2000]; local SSA performed in each window of width $W = 3M$, with $M = 85$ yr.

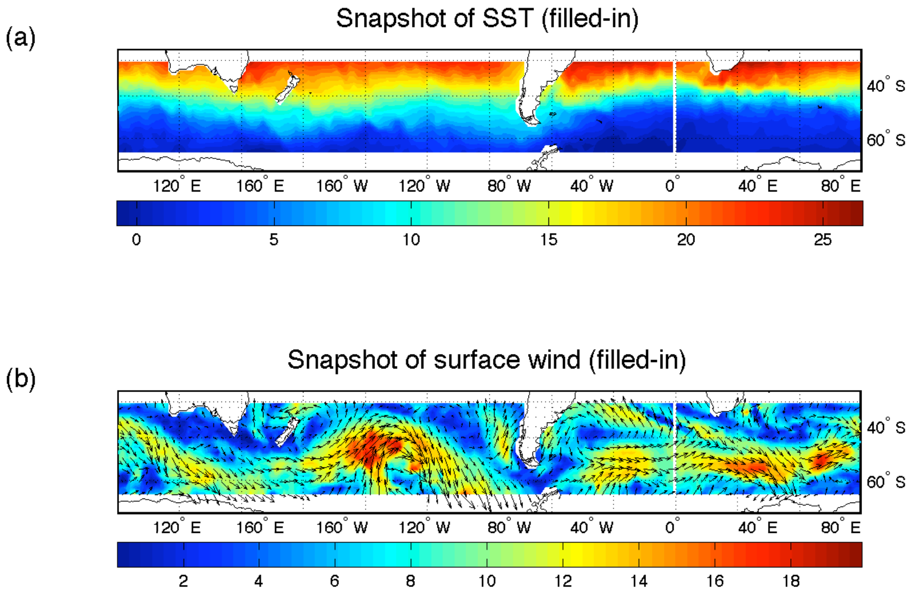
Radiation belts: synthetic gaps

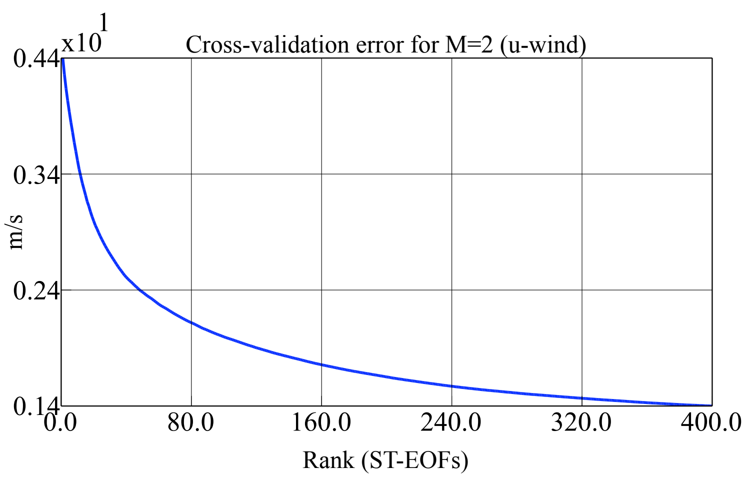


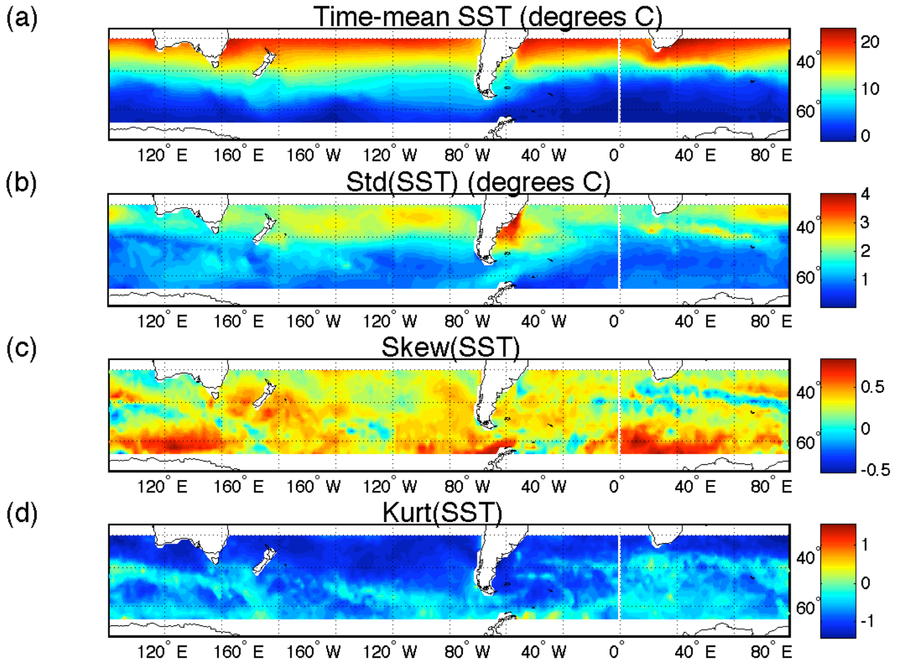


- Large gaps for different storms are filled-in.

Filled-in Southern Ocean data

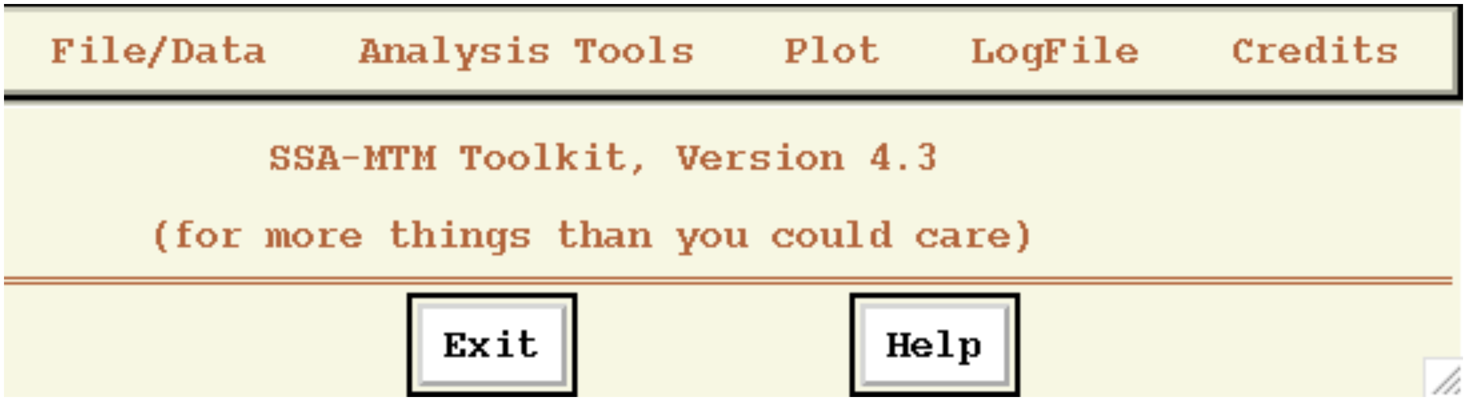






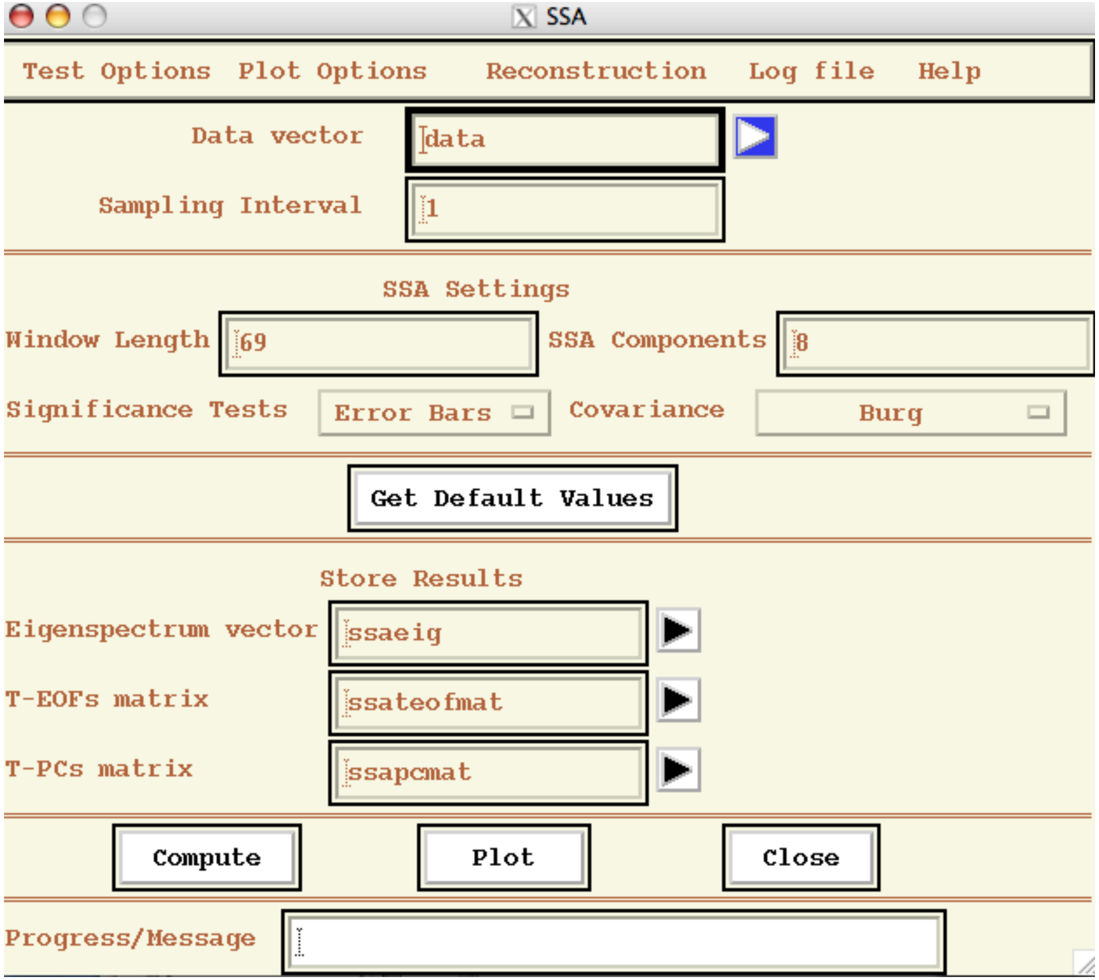
- Gap-filling needs to respect physical limits
- Complete dataset with full statistics indicates important nonlinear features.

SSA-MTM Toolkit



- Freeware ported to Sun, Dec, SGI, PC Linux, and Mac OS X
- Graphics support for IDL and Grace (free)
- Includes ***Blackman-Tukey FFT***, ***Maximum Entropy Method***, ***Multi-Taper Method (MTM)***, ***SSA and M-SSA***.
- Spectral estimation, decomposition, reconstruction & prediction.
- Significance tests of "**oscillatory modes**" vs. "**noise.**"
- Gap-filling coming shortly.

SSA-MTM Toolkit (cont'd)



- Data management with ***named vectors & matrices.***
- ***Default values.***
- Precompiled binaries are available at www.atmos.ucla.edu/ tcd/ssa

Selected References

- *Ghil M., R. M. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, M. E. Mann, A. Robertson, A. Saunders, Y. Tian, F. Varadi, and P. Yiou, 2002: "Advanced spectral methods for climatic time series," Rev. Geophys., 40(1), pp. 3.1-3.41, 10.1029/2000RG000092.*
- *D. Kondrashov and M. Ghil, 2006: Spatio-temporal filling of missing points in geophysical data sets,
Nonl. Proc. Geophys., 13, 151-159.*
- *more at http://www.atmos.ucla.edu/tcd/ssa*