

Gap filling in incomplete geophysical data sets

D. Kondrashov, M. Ghil¹

Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles
 1. Additional affiliation: Département de Géosciences, Ecole Normale Supérieure, Paris.



Introduction

Data sets in the geosciences are often full of gaps. This is usually the case for geological and paleoclimatological proxy data from the remote past, as well as for historical records from the more recent past. In modern times, data points may be missing because of the way the measurements are obtained.. We demonstrate here how Singular Spectrum Analysis (SSA) and multi-channel SSA can be applied to fill the missing data with smooth information from an iteratively inferred “signal” that represents coherent spatio-temporal structures, while the “noise” variance is discarded or reduced.

Gap filling by iterative SSA

1. Choose window M and start with $K=1$. Flag fraction of dataset $X(t)(t=1:N)$ as “missing” for cross-validation.

2. Update $X(t)$ mean and apply SSA embedding procedure for $N' = N - M + 1$:

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

3. Update covariance, find leading K EOFs

$$C_X = \frac{1}{N'} D^t D; C_X E_k = \lambda_k E_k$$

4. Reconstruct missing points using K EOFs

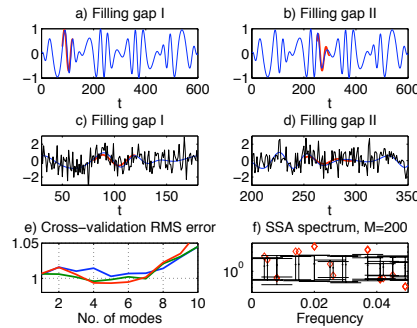
$$A_k(t) = \sum_{j=1}^M X(t+j-1) E_k(j)$$

$$R_X(t) = \frac{1}{M'} \sum_{k=K} \sum_{j=L_t}^{U_t} A_k(t-j+1) E_k(j);$$

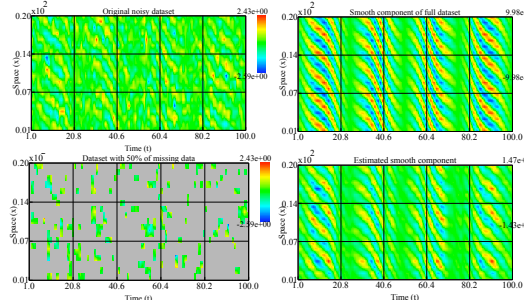
5. Check cross-validation error: if convergence has occurred for the missing points, set $K = K + 1$; if not, go back to Step 2.

- For a window width $M > 1$, one utilizes both spatial and temporal correlations in a multivariate data set.
- The optimal number K of “signal” EOFs is obtained through cross-validation.

Synthetic Examples

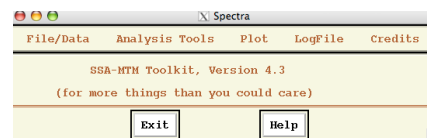


Gap filling of (a,b) a nonlinear oscillatory signal $s(t)$ (blue line), and (c,d) of its noise-contaminated version $x(t)$ (black line); red line is the filled-in data. (e) Cross-validation tests for $M = 160, 180$ and 200 .



Gap filling of a multivariate oscillatory spatio-temporal pattern contaminated by noise.

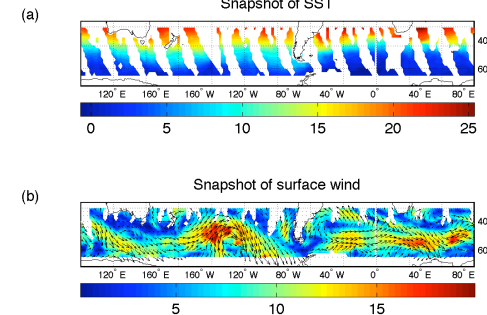
SSA-MTM Toolkit



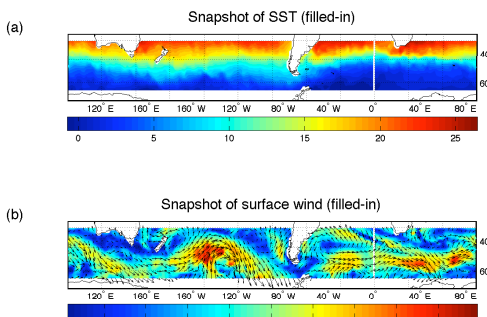
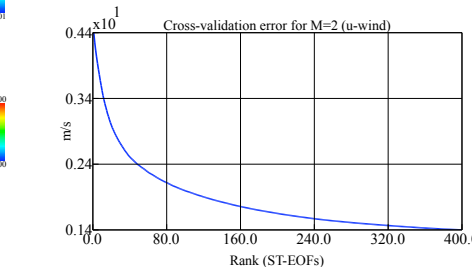
- Freeware for time series analysis.
- X11 (Motif) GUI.
- Ported to Sun, Linux, and Mac OS X.
- Downloads at <http://www.atmos.ucla.edu/tcd/ssa/>

Southern Ocean Data

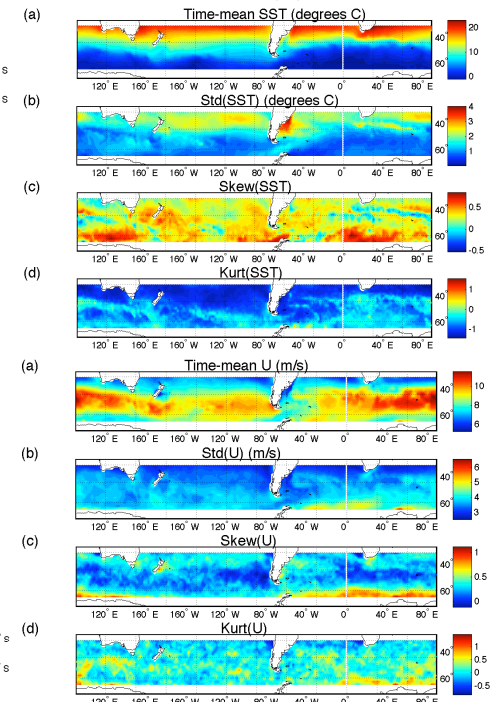
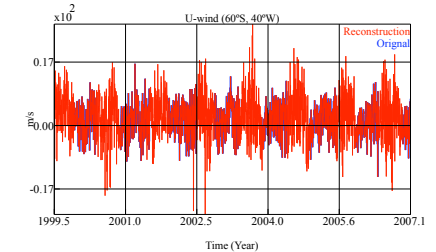
SST (AMSR-E), daily $2^\circ \times 2^\circ$, June 2002 – February 2007, 38.2% of missing points.
 Wind (QuikSCAT), daily $2^\circ \times 2^\circ$, July 1999 – February 2007, 17.2% of missing points.



Gap filling with $M = 5$ for SST and $M = 2$ for winds



Filled-in set



References

1. Ghil M., R. M. Allen, M. D. Dettinger, K. Ide, D. Kondrashov et al., 2002: Advanced spectral methods for climatic time series, *Rev. Geophys.*, 40(1), pp. 3.1-3.41, 10.1029/2000RG000092
2. D. Kondrashov and M. Ghil, 2006: Spatio-temporal filling of missing points in geophysical data sets, *Nonlin. Proc. Geophys.*, 13, 151--159