A BIVARIATE EXTENSION TO TRADITIONAL EMPIRICAL ORTHOGONAL FUNCTION ANALYSIS

ALLAN AASBJERG NIELSEN AND KLAUS BAGGESEN HILGER

Informatics and Mathematical Modelling, Technical University of Denmark, Building 321
DK-2800 Kgs. Lyngby, Denmark, Telephone +45 4525 3425, Telefax +45 4588 1397
{aa,kbh}@imm.dtu.dk, www.imm.dtu.dk/~{aa,kbh}

OLE BALTAZAR ANDERSEN AND PER KNUDSEN

National Survey and Cadastre, Rentemestervej 8
DK-2400 Copenhagen NV, Denmark, Telephone +45 3587 5050, Telefax +45 3587 5051
{oa,pk}@kms.dk, research.kms.dk/~{oa,pk}

This paper describes the application of canonical correlations analysis to the joint analysis of global monthly mean values of 1996-1997 sea surface temperature (SST) and height (SSH) data. The SST data are considered as one set and the SSH data as another set of multivariate observations, both with 24 variables. This type of analysis can be considered as an extension of traditional empirical orthogonal function (EOF) analysis which provides a marginal analysis of one variable over time. The motivation for using a bivariate extention stems from the fact that the two fields are interrelated as for example an increase in the SST will lead to an increase in the SSH. The analysis clearly shows the build-up of one of the largest El Niño events on record. Also the analysis indicates a phase lag of approximately one month between the SST and SSH fields.

1 Introduction

This paper deals with an extension of the method of empirical orthogonal functions (EOF) [14]. EOF analysis is often used in oceanography and other geophysical sciences to analyse temporal sequences of scalar (image) data. In this paper we extend this type of analysis by applying canonical correlation analysis to two temporal sequences of scalar image data, namely global sea surface temperature (SST) and global sea surface height (SSH). This type of analysis can be extended further to more than two sets of data [6,10,11]. The data used to illustrate the analysis carried out come from the Ocean Pathfinder programmes set up by NASA/NOAA, [8,13].

The data chosen represent relevant oceanographic properties related to one of the largest El Niño events ever recorded. El Niño is a large-scale warm ocean event in the Pacific off the coast of Peru and Ecuador caused by eastward drifting toward the west coast of South America of the pool of warm waters normally residing in the western part of the Pacific. This event is not local but may influence weather conditions worldwide with potentially drastic socio-economic consequences.

2   Method

EOF analysis is a name often used in geophysical data processing for principal component analysis (PCA) [4,1]. Often the usual PCA assumption on variables with mean zero is replaced by an assumption of temporal means of zero. PCA finds linear combinations (called principal components, PCs) of the original data that maximise variance. PC1 contains maximum variance. Higher order PCs contain maximum variance subject to orthogonality or uncorrelatedness with lower order PCs. Canonical correlation analysis (CCA) [5,2,1] jointly analyses two sets of variables. It finds two sets of linear combinations (called canonical variates, CVs) of the zero mean original variables that maximise correlation between the two. The two CV1s are linear combinations of the original data (one from each set) that are maximally correlated. Higher order CVs are maximally correlated subject to orthogonality or uncorrelatedness with lower order CVs. The correlations obtained between corresponding CVs are termed canonical correlations.

The desired linear combinations $a^T X$ and $b^T Y$ of the two sets of variables X and Y are obtained by maximising the correlation

$$\rho = a^T \Sigma_{12} b \Big/ \sqrt{(a^T \Sigma_{11} a)(b^T \Sigma_{22} b)}$$

between the two. Σ_{11} and Σ_{22} are dispersion matrices of X and Y, respectively, and Σ_{12} is the covariance between them.

The desired linear combinations are obtained by solving the generalised eigenvalue problems

$$\Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} a = \rho^2 \Sigma_{11} a$$
$$\Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} b = \rho^2 \Sigma_{22} b$$

where a and b are (conjugate eigen-) vectors of the desired weights ($\Sigma_{21} = \Sigma_{12}^T$). The maximised ρ is called the canonical correlation; the corresponding $a^T X$ and $b^T Y$ are the canonical variates.

3   Data and results

The data used are global monthly mean values of 1996-1997 SST data from the NOAA/NASA Oceans Pathfinder AVHRR SST database, [13], and global monthly mean values of 1996-1997 SSH data from the NASA/GSFC Ocean Altimeter Pathfinder database, [8]. The SSH data are interpolated point observations from the TOPEX/Poseidon radar altimeter mission. The SST data come as 360 rows by 720

columns half degree data starting at 180° longitude, the SSH data come as 179 rows by 360 columns one degree data starting at 0° longitude. The SST data have been resampled to the SSH grid. The AVHRR instrument is influenced by cloud coverage whereas the radar altimeter provides uninterrupted data. Statistics for the CCA are calculated only where both variables have non-missing values for all 24 months.

Figure 1 shows the first six SST CVs resulting from an analysis where the 24 months of SST data are considered as one set and the 24 months of SSH data are considered as the other set. The CVs are stretched linearly from mean ± three standard deviations, the pseudo-colour scale goes from blue (minimum) over cyan, green, yellow to red (maximum).

Figure 2 shows the first six SSH CVs (same analysis, stretch and colouring).

Figure 3 shows the canonical correlations, and the correlations between the original data and the first three CVs for both SST and SSH.
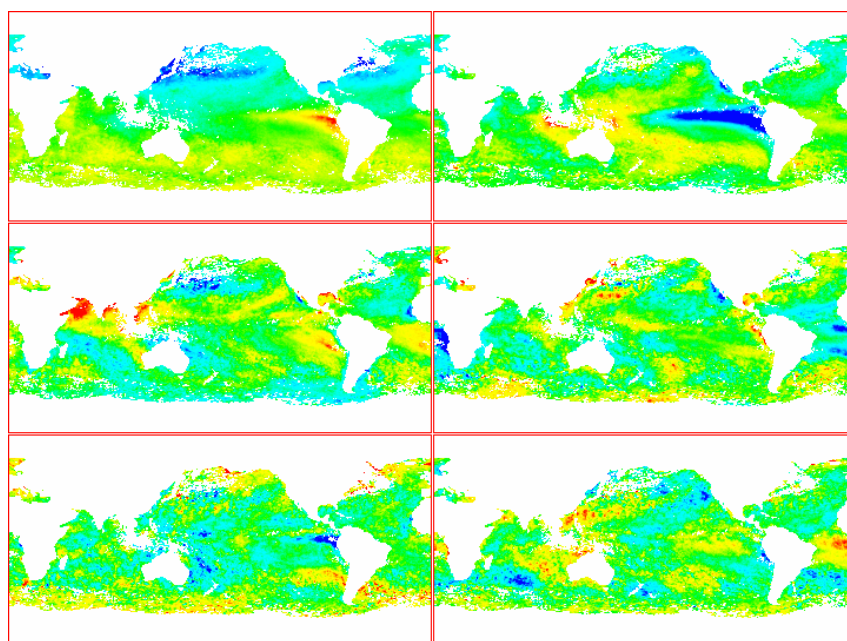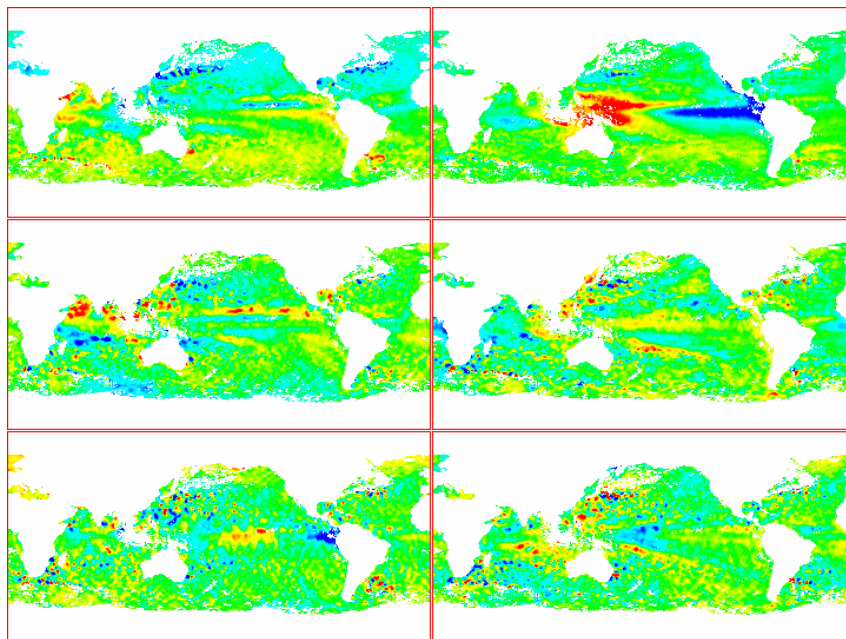

Figure 1. SST CVs 1-6 row-wise.

Figure 2. SSH CVs 1-6 row-wise.

4 Discussion and conclusions

The first pair of CVs has a correlation coefficient of 0.7587 and accordingly the two exhibit very similar spatial patterns. Generally speaking, we see highs in the Southern Hemisphere and lows in the Northern Hemisphere. This is in accordance with the observed maximally negative correlations between both SST and SSH and both CV1s in the Northern Hemisphere's summer months. A high in SSH CV1 and a strong high in SST CV1 off the equatorial west coast of South America is seen. This is in accordance with the El Niño phenomenon. The conspicuous oscillation is clearly at an annual period, but the correlations between CV1 and SST remain negative and do not alternate around zero which should be expected. The highest correlations occur in month 9 and 21 for both the SST and the SSH. Normally the highest correlation for the SST would occur one month earlier, but because of the joint analysis of the SST and SSH, this is dominated by the SSH. This suggests an analysis of SST and a temporally shifted version of SSH.
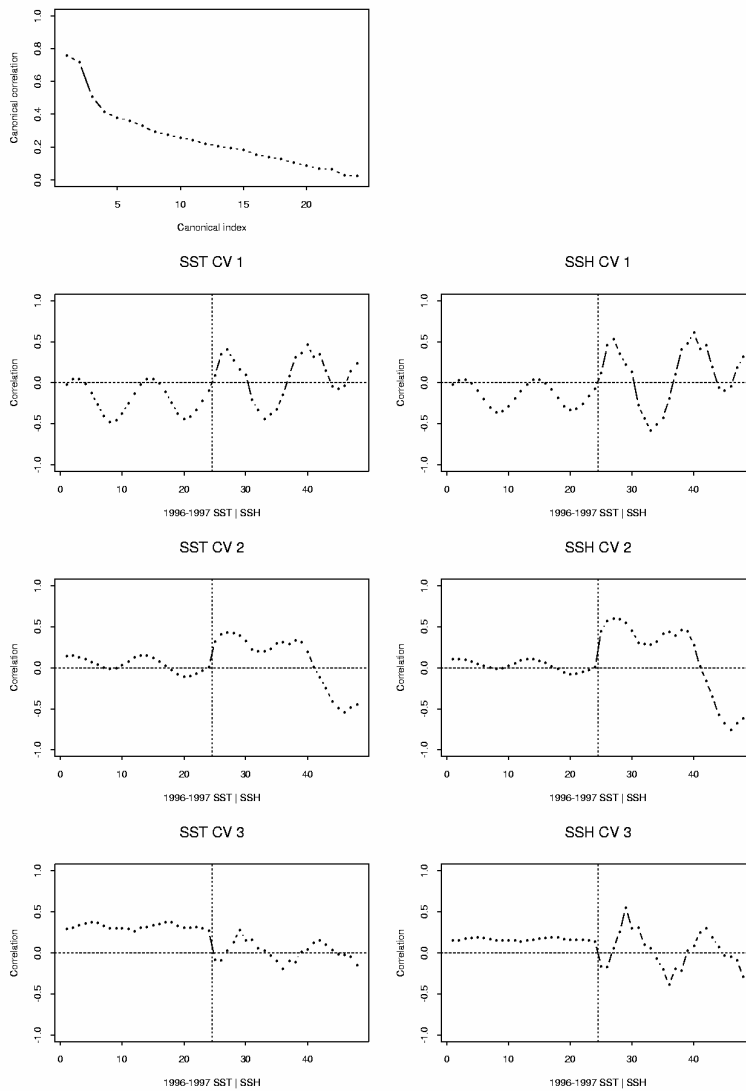
Figure 3. Canonical correlations (top), and correlations between original SST (indices 1-24, left of the dashed vertical line) and SSH variables (indices 25-48, right of the dashed vertical line), and SST CVs 1-3 (left column) and SSH CVs 1-3 (right column).

The second pair of CVs has a correlation coefficient of 0.7179 and again the two exhibit similar spatial patterns. For this CV pair we see strong lows in both SSH CV2 and SST CV2 off the equatorial west coast of South America which again is in accordance with the El Niño event. Also, a high in SST CV2 and a very strong high in SSH CV2 east of Indonesia and the Philippines is seen.

The correlations between the original data and the CVs indicate that both CVs 1 and CVs 2 reflect strong annual oscillations [7]. Especially for SSH and CVs 2 but also for SSH and CVs1 and to a lesser degree for SST and CVs2 these oscillations seem to be disturbed in the second half of 1997. This is agreement with the spatial patterns observed in especially CVs 2 and the El Niño build-up during the last eight months of 1997.

The third canonical correlation is 0.5061, and the fourth is 0.4148. This is still relatively high, but the fact that the first two canonical correlations are much higher than the remaining ones shows that the joint variation of the two geophysical fields over the two-year period is well explained by only two CVs. On the other hand the two highest canonical correlations are only 0.7587 and 0.7179, and therefore there are also variations in the two fields that are not well explained by a joint (linear) analysis. For a nonlinear analysis of the same data, see [3].

In spite of the very short time span covered by the data and the associated risk of over-interpreting the results, simultaneous inspection of spatial patterns of the CVs and the correlations between the original and transformed variables from the analysis gives good indications of an anomaly off the South American west coast taking place in the second half of 1997. This is in good agreement with established oceanographic knowledge on the build-up of one of the largest El Niño events on record.

Future analysis should include longer time series in order to establish whether 1997 (and 1998) really represent anomalous events in terms of global SST and SSH.

Investigations should also be made in which the SST field is shifted temporally to the SSH field in order to investigate if the correlations between the CV and the individual SST and SSH fields could be increased. The apparent phase lag of one month between the highs of the SST and the SSH is explained by the fact that the SST represent the instant temperature of the sea surface, whereas the steric expansion causing the sea level to rise is a more integrated effect.

Acknowledgement

References

1. Anderson T. W., *An Introduction to Multivariate Statistical Analysis*, second edition, John Wiley, 1984.
2. Cooley W. W. and Lohnes P. R., *Multivariate Data Analysis*, John Wiley and Sons, 1971.
3. Hilger K. B., Nielsen A. A., O. B. Andersen and Knudsen, P., An ACE-based nonlinear extension to traditional empirical orthogonal function analysis, *MultiTemp 2001*.
4. Hotelling H., Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, 24, 417-441, 1933.
5. Hotelling H., Relations between two sets of variates, *Biometrika*, XXVIII, 321-377, 1936.
6. Kettenring, J. R., Canonical analysis of several sets of variables, *Biometrika*, 58, 433-451.
7. Knudsen P., Andersen O. B. and Knudsen T., ATSR sea surface temperature data in a global analysis with TOPEX/Poseidon altimetry, *Geophysical Research Letters,* 23(8), 821-824, 1996.
8. NASA/GSCF, http://neptune.gsfc.nasa.gov/ocean.html, Goddard Space Flight Center, National Aeronautics and Space Administration, Greenbelt, Maryland, USA.
9. NASA Facts, *El Niño*, The Earth Science Enterprise Series, NF-211, 1999.
10. Nielsen, A. A., *Analysis of regularly and irregularly sampled spatial, multivariate, and multi-temporal data*, Ph.D. thesis, IMM-, Technical University of Denmark, http://www.imm.dtu.dk/~aa/phd, 1994.
11. Nielsen A. A., Multiset canonical correlations analysis and multispectral, truly multi-temporal remote sensing data, *IEEE Transactions on Image Processing* accepted 2001.
12. Nielsen. A. A., Hilger K. B., Andersen O. B. and Knudsen P., A temporal extension to traditional empirical orthogonal function analysis, *MultiTemp 2001*.
13. PO.DAAC, http://podaac.jpl.nasa.gov, Jet Propulsion Laboratory, National Aeronautics and Space Administration, Pasadena, California, USA.
14. Preisendorfer R. W., *Principal Component Analysis in Meteorology and Oceanography*, posthumously compiled and edited by C. D. Mobley. Developments in Atmospheric Science, 17, Elsevier, 1988.