

Rotation of EOFs by the Independent Component Analysis: Toward a Solution of the Mixing Problem in the Decomposition of Geophysical Time Series

FILIFE AIRES

NASA Goddard Institute for Space Studies, Department of Applied Physics, Columbia University, New York, New York

WILLIAM B. ROSSOW

NASA Goddard Institute for Space Studies, New York, New York

ALAIN CHÉDIN

CNRS Laboratoire de Météorologie Dynamique, Palaiseau, France

(Manuscript received 11 January 2001, in final form 14 May 2001)

ABSTRACT

The Independent Component Analysis (ICA) is a recently developed technique for component extraction. This new method requires the statistical independence of the extracted components—a stronger constraint that uses higher-order statistics—instead of the classical decorrelation (in the sense of “no correlation”), which is a weaker constraint that uses only second-order statistics. This technique has been used recently for the analysis of geophysical time series with the goal of investigating the causes of variability in observed data (i.e., exploratory approach). The authors demonstrate with a data simulation experiment that, if initialized with a Principal Component Analysis (PCA), the ICA performs a rotation of the classical PCA (or EOF) solution. This experiment is conducted using a synthetic dataset, where the correct answer is known, to more clearly illustrate and understand the behavior of the more familiar PCA and less familiar ICA. This rotation uses no localization criterion like other rotation techniques; only the generalization of decorrelation into full statistical independence is used. This rotation of the PCA solution seems to be able to avoid the tendency of PCA to mix several components, even when the signal is just their linear sum.

1. Introduction

This work concerns methods for the investigation of the physical causes of variability of a dynamical system, for example, the climate, from observations of its behavior. The observed time series of the system's state can be produced by a mixture of different components representing different physical processes. In the most general case, the time series $\mathbf{x}(j)$ with temporal dimension N at a particular spatial coordinate $j \in \{1, \dots, M\}$, where M is the spatial dimension, is the result of the mixture of these components $\boldsymbol{\sigma}(j)$ by an operator \mathcal{G} :

$$\mathbf{x}(j) = \mathcal{G}[\boldsymbol{\sigma}(j)]. \quad (1)$$

In this paper we use lower (upper) case bold letters to indicate vectors (matrixes) and we will refer to a particular spatial location as a pixel. We consider decomposition in time—that is, the observations, $\mathbf{x}(j)$, are time

series at each pixel, j —but the following discussion would be the same for a decomposition in space.

The goal of the analysis method is to infer the unknown contributing components $\boldsymbol{\sigma}(j)$ from the observed data $\mathbf{x}(j)$. Often in performing such an analysis, one does not know a priori much about what \mathcal{G} is like, even whether it is linear or not. We introduce

$$\mathbf{h} = \mathcal{H}(\mathbf{x}) \approx \boldsymbol{\sigma}, \quad (2)$$

where \mathcal{H} is an estimate of the unknown inverse mapping \mathcal{G}^{-1} and \mathbf{h} is an estimate of the unknown vector $\boldsymbol{\sigma}$. Analysis methods that estimate \mathcal{H} and \mathbf{h} are called component extraction techniques.

If \mathcal{G} is nonlinear and cannot be usefully linearized, then we are faced with a component extraction problem that is highly complex for many reasons. First, the definition of a well-adapted nonlinear component extraction model \mathcal{H} is difficult without some a priori information about the nonlinear mixture model \mathcal{G} . Using generic statistical models for the nonlinear regressions of \mathcal{H} often introduces too many degrees of freedom, which ruins the inference process. Second, the determination of the extracted components is much more complex

Corresponding author address: Dr. Filipe Aires, NASA Goddard Institute for Space Studies, Columbia University, 2880 Broadway and 112th Street, New York, NY 10025.
E-mail: faires@giss.nasa.gov

since the basis functions $\mathbf{g}_i[\boldsymbol{\sigma}(j)]$, vary with location j . Third, the uniqueness of the result and its interpretation in physical-process terms is difficult to demonstrate. Some nonlinear techniques have been developed (Monahan 2000) but their main application is to represent complex data in a more compact way (i.e., compression). Here, we are not concerned with this use of such techniques, but with the extraction of “meaningful” (i.e., causal) components to understand how a dynamical system works. Currently, all the “classical” (and most frequently used) extraction methods are linear; the use of nonlinear models in (2) for component extraction is only just beginning. The nonlinear case is beyond the scope of this paper.

One approach that may be useful when \mathcal{G} is complex is to linearize (1) using $\mathbf{G}[\boldsymbol{\sigma}^0(j)]$, the Jacobian matrix of the nonlinear operator \mathcal{G} near a particular state, $\boldsymbol{\sigma}^0(j)$:

$$\begin{aligned}\Delta \mathbf{x}(j) &= \mathbf{G}[\boldsymbol{\sigma}^0(j)] \cdot \Delta \boldsymbol{\sigma}(j) \\ &= \mathbf{g}_1[\boldsymbol{\sigma}^0(j)] \cdot \Delta \sigma_1(j) + \mathbf{g}_2[\boldsymbol{\sigma}^0(j)] \cdot \Delta \sigma_2(j) \\ &\quad + \dots + \mathbf{g}_\varrho[\boldsymbol{\sigma}^0(j)] \cdot \Delta \sigma_\varrho(j),\end{aligned}\quad (3)$$

where Δ is the difference operator, and the temporal basis functions $\mathbf{g}_1[\boldsymbol{\sigma}^0(j)], \dots, \mathbf{g}_\varrho[\boldsymbol{\sigma}^0(j)]$, which are the columns of the matrix $\mathbf{G}[\boldsymbol{\sigma}^0(j)]$, are unknown time series describing a fixed dynamical behavior at each pixel, j . Each $\mathbf{g}_i[\boldsymbol{\sigma}^0(j)]$ is the temporal response to a perturbation $\Delta \sigma_i(j)$ of i th component at pixel j when the state is given by $\boldsymbol{\sigma}^0(j)$. For example, in the case where the physical component i is an oscillating wave propagating in space, the time series $\mathbf{g}_i[\boldsymbol{\sigma}^0(j)]$ have the same shape as the source but with a time delay dependent on the distance between their location at pixel j and the source of the wave.

When \mathcal{G} is known a priori to be linear or has been linearized to \mathbf{G} , the equivalent to (1)–(4) when the time series $\mathbf{x}(j)$ at particular spatial coordinate $j \in \{1, \dots, M\}$ is decomposed in time, is:

$$\begin{aligned}\mathbf{x}(j) &= \mathbf{G} \cdot \boldsymbol{\sigma}(j) \\ &= \mathbf{g}_1 \sigma_1(j) + \mathbf{g}_2 \sigma_2(j) + \dots + \mathbf{g}_\varrho \sigma_\varrho(j),\end{aligned}\quad (5)$$

where the temporal basis functions, $\mathbf{g}_1, \dots, \mathbf{g}_\varrho$, which are the columns of matrix \mathbf{G} , are unknown time series describing a fixed dynamical behavior. In contrast with the nonlinear case, the basis functions \mathbf{g}_i are independent of the geographical location j . Each \mathbf{g}_i could be a signal with a different physical cause operative in a particular geographical region represented by a different component map $\{\sigma_i(j); j = 1, \dots, M\}$. The goal of the analysis is to infer the unknown contributing components $\boldsymbol{\sigma}(j)$ from the observed data $\mathbf{x}(j)$. In the linear case, we write (2) as

$$\mathbf{h} = \mathbf{J} \cdot \mathbf{x} \simeq \boldsymbol{\sigma}, \quad (6)$$

where \mathbf{J} is an estimate of the unknown matrix \mathbf{G}^{-1} (the superscript $^{-1}$ represents the pseudoinverse if \mathbf{G} is not

square) and \mathbf{h} is an estimate of the unknown vector $\boldsymbol{\sigma}$. The ability of statistical analysis techniques to retrieve good estimates, \mathbf{h} , of the true components, $\boldsymbol{\sigma}$, is highly dependent on the quality of the statistical dataset used (i.e., a sufficiently large number of independent examples is needed to sample all the variations involved) and on the technical assumptions that are made about \mathbf{J} and \mathbf{h} .

There are many techniques that have been developed to estimate \mathbf{J} and \mathbf{h} . The one most frequently used by the climate research community today is the Principal Components Analysis (PCA) or empirical orthogonal function (EOF) method (Lorenz 1951; von Storch and Frankignoul 1998). Sometimes, modifications of this method are used that either apply some other criterion besides maximizing the variance explained by each component or relax the requirement for orthogonal basis functions. These methods use this additional information to perform an orthogonal or oblique rotation of a previous PCA solution; we refer here to this large set of methods as rotational techniques (RT; Horel 1981; Richman 1986). We formulate a general linear case and the classical analysis techniques in section 2.

The PCA solution has some well-known difficulties. 1) The spatial orthogonality of the eigenvectors is not well-adapted to all problems. As a consequence, many EOF solutions display an artificial alternating sign structure. Horel (1981) quotes many cases in the literature of this artificial feature, for example for the sea level pressure in restricted regions such as Australia or the Arctic. 2) The principal components depend on the domain used for the analyses (see, e.g., different results obtained for the EOF analysis of the sea level pressure in the North American region; Buell 1975). 3) Weighting of geophysical data in the EOF analysis and in rotated analysis seems to have an important impact on the results obtained (Chung and Nugam 1999). Furthermore, (Karl et al. 1982) comment on possible distortions when using irregularly spaced data. 4) EOF components can be hard to interpret physically because of all these problems. All these concerns have led to the development of the rotational techniques (Horel 1981; Richman 1986).

The Independent Component Analysis (ICA) is described in section 3. This method, which can be interpreted as a rotation technique, is based on information theory and has been recently developed in the context of signal processing studies and of the development of neural coding models (Jutten and Herault 1991; Atick 1992; Bell and Sejnowski 1995). This technique has now been studied for some time by the statistical analysis research community and many recent applications of the ICA paradigm can be found in the ICA 1999 proceedings (Cardoso et al. 1999) or Hyvärinen and Oja (2000), but this method has not been used for analysis of climatological observations (Aires et al. 2000). The two major distinctions between the ICA approach and the classical techniques are the following.

- The method extracts statistically independent components, even if these components have non-Gaussian probability distribution functions, by making use of higher-order statistics, whereas the PCA and RT approaches use only second-order statistics.
- A linear mixture model is not assumed; any extraction model could be used with the ICA paradigm (Burel 1992), which allows for the introduction of pertinent a priori information about the mixture model, if it is available.

We will show that the “linear” (this term will be explained in the following) and PCA-initialized form of ICA, described here, performs a rotation of the PCA solution, eliminating the mixing problem: PCA has the tendency to mix several components, even when the signal is just the linear sum. We argue that, because of certain general features of the ICA approach, it is a particularly promising technique for rotations of the PCA solution. Furthermore, a previous study that applies a similar ICA to the analysis of variations of tropical sea surface temperature time series (Aires et al. 2000) illustrates its potential to separate a geophysical time series into more meaningful components.

To illustrate most clearly how ICA avoids the component mixing problem, we construct a synthetic dataset, where the true answer to the decomposition problem is known, and apply PCA-initialized Independent Component Analysis to extract components (section 4). We have deliberately devised a dataset that exaggerates characteristics found in the real atmospheric observations to, on the one hand, make it even easier for PCA to separate the components and to, on the other hand, challenge whether ICA can separate distinct modes of variability when they are similar in their space and time distributions. In particular, the synthetic dataset is formed by a linear sum of components, some of which are better-separated in space and time than in real atmosphere, and some of which overlap in space, are correlated in time, or represent teleconnections as in the real case. Thus, this dataset is created so that it has structures that a linear component extraction techniques should find. If a method fails on this simple case, it is unlikely to succeed when applied to the more ambiguous climate case. We show that, even in the simple linear case, the PCA technique mixes the components incorrectly. We also show that the ICA method performs a rotation to correctly separate the components, but, as for all statistical component extraction methods, this is only a necessary condition not a sufficient guarantee of success for climate analysis.

The goals of this paper are then to illustrate some of the problems of the most commonly used classical analysis technique, even in the simple linear case, and to introduce a new component extraction technique that overcomes these problems, at least in its linear form (section 5). We compare linear ICA to PCA to measure the effect of the rotation transformation by the ICA

algorithm. We do not use a priori information for the component extraction experiments since we are interested in exploratory techniques, not confirmatory ones; that is, we want a technique to find the correct but unknown components, not confirm results from another analysis.

2. The linear case and classical component extraction techniques

A common approach for statistical component extraction is to require the decorrelation of the extracted components, in which case the covariance matrix of extracted components $\langle \mathbf{h}' \cdot \mathbf{h} \rangle$ is constrained to be diagonal; but this decorrelation constraint has an infinity of solutions because

$$\mathbf{J} = \Theta \cdot \mathbf{J}_0, \quad (7)$$

where Θ is any undetermined $Q \times Q$ matrix so that $\Theta' \cdot \Theta = \mathbf{I}_{Q \times Q} \cdot \mathbf{J}_0 = \Sigma^{-1/2} \cdot \mathbf{E}'$ is a $Q \times N$ matrix with Σ the truncated diagonal matrix of the larger (in decreasing order) eigenvalues of $\langle \mathbf{x}' \cdot \mathbf{x} \rangle$ and \mathbf{E} the $N \times Q$ matrix with the associated normalized eigenvectors in the columns.

One particular decorrelation solution is the well-known PCA or, in the geophysical community, EOF¹ analysis, first used in atmospheric sciences by Lorenz (1951). In this technique, an additional constraint is added to resolve the indeterminacy of the decorrelation solutions: successive extracted components have to explain the maximum remaining variance. This solution is given by taking $\Theta = \mathbf{I}_{Q \times Q}$ in (7). Depending on which space the PCA is applied to (space, time, frequency, multivariate data, etc), the PCA has also been called Singular Spectrum Analysis (Broomhead and King 1986; Vautard et al. 1992), Multichannel Singular Spectrum Analysis (Vautard et al. 1996), Extended Empirical Orthogonal Functions (Korres et al. 2000), Multivariate Empirical Orthogonal Function (Xue et al. 2000), etc.

Three well-known problems arise when using the PCA technique. 1) Even if the mixing of the components is linear as in Eq. (5), the maximum-explained-variance assumption can lead to a different mixing in the extracted components (Kim and Wu 1999) as we will be shown here (see Fig. 1a for an schematic illustration of this problem in a two-dimensional case). 2) This mixing problem is also particularly serious when the PCA is applied to data that have more than one component with about the same variance. In this case, the problem is not solvable since any orthogonal rotation of the principal components (i.e., in the space of the “degenerate” eigenvectors) will be a PCA solution (Fig. 1b). 3) Since PCA imposes orthogonality on the extracted basis functions, mixing problems also arise when the actual physical basis functions are not orthogonal (Fig. 1c). Another

¹ EOF is a specific form of the general PCA where the extracted basis functions are normalized.

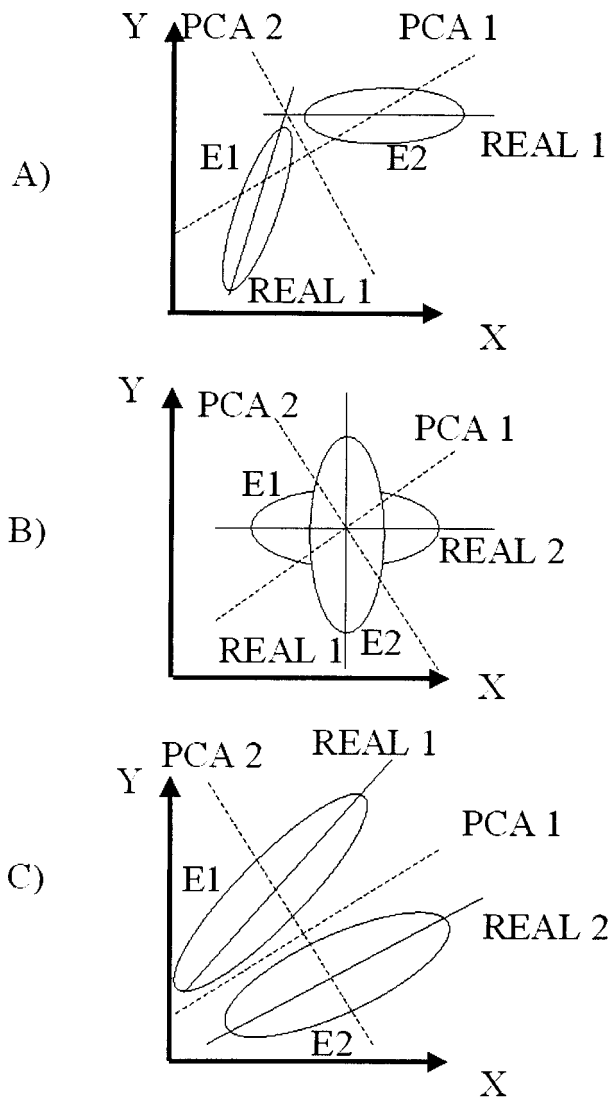


FIG. 1. Illustration of the problems encountered by PCA when observations, two-dimensional (coordinates x and y), come from two components defining ellipses E1 and E2. The line D represents the first PCA axis defining the first PCA component: (a) mixing due to the maximum-explained-variance constraint, (b) indeterminacy when two components have same variance, and (c) mixing due to the non-orthogonality of components.

problem for the application of the PCA to geophysical data arises from an irregularly spaced grid of pixels that can lead to distorted basis functions (Karl et al. 1982).

The PCA assumptions (linearity, orthogonality, maximum variance explained by each successive component) used to resolve the solution indeterminacy are not known, a priori, to be valid for a particular dataset. They are not likely to be valid for climate variations. If these assumptions are not valid, variations that are not physically connected could be artificially mixed together into one extracted component (i.e., the mixing problem). This is the reason why PCA is often used in restricted geographical domains instead of global domains or ap-

plied to prefiltered data to try to isolate a single dominant mode of variation, which PCA can correctly identify. Thus, although PCA is useful for compressing information by *describing* the most variance with the fewest terms in an expansion (as a dimension-reduction/compression technique), it can lead to misinterpretation of physical relationships when used as a component extraction technique.

Rotational techniques were introduced (Horel 1981; Richman 1981), in part, to obtain a more physically interpretable solution and to avoid some of the problems of PCA. In these approaches, an additional constraint of localization, based on the so-called simple structure principle, is used to solve the indeterminacy of the decorrelation solutions. The rotations are said to be orthogonal (the rotation matrix is an orthogonal matrix) or oblique (this constraint is relaxed). There exist many proposed localization criteria: quartimax, varimax, transvarimax, quartimin, oblimax, etc. [see the review paper of Richman (1986) on this subject]. Two general distinct classes of RT solution can be distinguished: confirmatory RT where a priori information about the components is available and we want to verify the hypothesis, and exploratory RT where almost no a priori information about the problem is available. We are interested in the exploratory case where no a priori information is available. Since no general principle for choosing a particular localization criterion from this large set of proposed solutions is available, use of a particular RT method in exploratory mode may be equivalent to introducing a priori information about the localization that may not be any better suited to the particular problem than PCA.

The most commonly used criterion for orthogonal rotation is

$$V_{\gamma}(\mathbf{G}) = \sum_{j=1}^Q \sum_{i=1}^N (G_{ij})^4 - \frac{\gamma}{N} \sum_{j=1}^Q \left[\sum_{i=1}^N (G_{ij})^2 \right]^2, \quad (8)$$

where the constant γ gives a family of rotations with $\gamma = 1.0$ giving varimax rotations, and $\gamma = 0.0$ giving quartimax rotations. The implementation of these techniques is not trivial, but automatic routines are available (see, e.g., the routine G03BAF in the NAG FORTRAN Library Routine Document).

Despite the proposed alternatives to the variance maximization assumption and the orthogonality constraint used in various RT methods, they still all share two fundamental properties with PCA: they assume that the meaningful components are linearly mixed (classical techniques are intimately linked to the linear assumption and cannot be generalized to nonlinear models) and that only second-order statistics need be evaluated.

3. The Independent Component Analysis technique

In this section, we describe the main concepts underlying the ICA technique. For more details, the in-

terested reader is referred to Bell et al. (1995) and Aires et al. (2000). The ICA technique aims to extract statistically independent components, a stronger constraint than the decorrelation requirement of the PCA.

The statistical independence of two variables, h_1 and h_2 , is determined when their joint probability distribution can be factored:

$$P(h_1, h_2) = P(h_1) \cdot P(h_2). \quad (9)$$

This constraint involves higher-order statistics whereas the decorrelation constraint only involves second-order statistics (i.e., mean and variance). Decorrelation is equivalent to statistical independence only in the case where the quantities are Gaussian distributed, so the higher-order statistics are particularly important when the analyzed data have components with non-Gaussian distributions (Comon 1994). Avoiding the a priori assumption that second-order statistics are sufficient is important when the components are unknown as is usually the case, but especially when some components may be partly correlated (as we will show).

It is also important to distinguish the non-Gaussian character of the components, σ , from the non-Gaussian character of the data \mathbf{x} in Eq. (5). If the data have a non-Gaussian distribution, then at least one component is also non-Gaussian, since for the simplest linear mixture of Gaussian components (not to be confused with a "mixture of Gaussians," which usually means that the random variable has one of a number of possible distributions), the distribution would be Gaussian, but a nonlinear combination of Gaussian distributions could also be non-Gaussian. Some previous studies examine the non-Gaussian behavior of geophysical data (Burgers and Stephenson 1999; Aires et al. 2000).

A variable is characterized by all its statistical cumulants: the first cumulant is the mean, the second is the variance, the third is the skewness, the fourth cumulant is the kurtosis, etc. (Press et al. 1992). For Gaussian variables, cumulants higher than 2 are zero. When data have a zero mean, the "skewness" $\text{skew}(X) = \langle X^3 \rangle / \sigma^3$ and the "kurtosis" $\text{kurt}(X) = \langle X^4 \rangle / \sigma^4 - 3$. These cumulants are often used to test the departure from Gaussian behavior. The skewness measures the symmetry of the probability distribution function: when the skewness is positive, larger events are more probable than smaller events, and the reverse is true when the skewness is negative. The kurtosis is a measure of the sharpness of the distribution: a negative kurtosis indicates that the distribution has a broader central peak and larger tails than a Gaussian distribution (sub-Gaussian), a positive kurtosis indicates that the distribution has a sharper central peak (super-Gaussian distribution). The non-Gaussian character of a variable is intimately linked to nonlinear dynamics (Palmer 1999). For example, a nonlinear dynamical system with two attractors can result in bimodal distributions. Thus, without a priori information on the Gaussianity of components in an analysis of geophysical time series, the use of ICA is rec-

ommended since its requirement of statistical independence is more general than the decorrelation assumption.

The time series observations are gathered into a dataset, \mathbf{X}_t^j , of M observations, $\mathbf{x}(j) = (x_j^t; t = 1, \dots, N)$, with $j \in \{1, \dots, M\}$, where M is the spatial dimension of the time series and N is its temporal dimension. The time series $\mathbf{x}(j)$ is assumed to be a mixture of statistically independent components $\sigma = \{\sigma_i; i = 1, \dots, Q\}$:

$$\mathbf{x}(j) = \mathcal{G}[\sigma(j)] \quad (10)$$

where \mathcal{G} is an unknown mixture operator, which is, by hypothesis, nonsingular (i.e., it can be inverted).

The goal of ICA is to retrieve a function $\mathcal{J}: \mathbf{x} \rightarrow \mathbf{h}$, where \mathbf{h} is an estimate of σ and the terms $\{h_i; i = 1, \dots, Q\}$ are statistically independent. The estimate \mathbf{h} is defined as a deterministic function (linear or not) of the observations:

$$h_i = \mathcal{J}_i(\mathbf{W}_i, \mathbf{x}); \quad i = 1, \dots, Q, \quad (11)$$

where $\{\mathbf{W}_i; i = 1, \dots, Q\}$ is the set of parameters of \mathcal{J} . As in RT, the number of components, Q , is here supposed to be known. This number can be estimated, in easy cases, by a break in the frequency spectrum of the data; for more difficult spectra, see for example (Jolliffe 1986). With real observations, Q depends on the analysis objectives: extracting a lot of components allows for more complete description of the variability but makes the interpretation much more complicated, whereas extracting fewer components focuses attention on fewer different phenomena at the cost of explaining less of the variability. The reader interested in this topic should refer to an article by Nadal et al. (2000).

The parameters, \mathbf{W}_i , are estimated by applying a gradient descent algorithm to a cost function that specifies the statistical independence of the $\{h_i; i = 1, \dots, Q\}$. Different equivalent cost functions can be used; we use here the *infomax* approach to ICA (Nadal and Parga 1994) for which simple algorithms have been derived (Bell and Sejnowski 1995). Information theory is used to specify the statistical independence cost function: the fundamental quality is *information redundancy*. Given Q variables, h_1, h_2, \dots, h_Q , the information redundancy, $\mathcal{R}(h_1, h_2, \dots, h_Q)$, is defined as the Kullback divergence (Dacunha-Castelle and Duflo 1982) between the joint distribution, $P_h(h_1, h_2, \dots, h_Q)$, and the factorized distribution, $P_1(h_1) \cdot P_2(h_2), \dots, P_Q(h_Q)$:

$$\mathcal{R}(\mathbf{h}) = \int_{-\infty}^{+\infty} \prod_{i=1}^Q dh_i P_h(\mathbf{h}) \log \frac{P_h(\mathbf{h})}{\prod_{i=1}^Q P_i(h_i)}. \quad (12)$$

This information redundancy measures the difference between the joint and the factorized distribution: when the redundancy $\mathcal{R}(\mathbf{h}) = 0$, $P_h(\mathbf{h}) = \prod_{i=1}^Q P_i(h_i)$, which means, by the definition in Eq. (9), that the components of vector \mathbf{h} are statistically independent. An important remark is that, since no geometric constraint on the basis

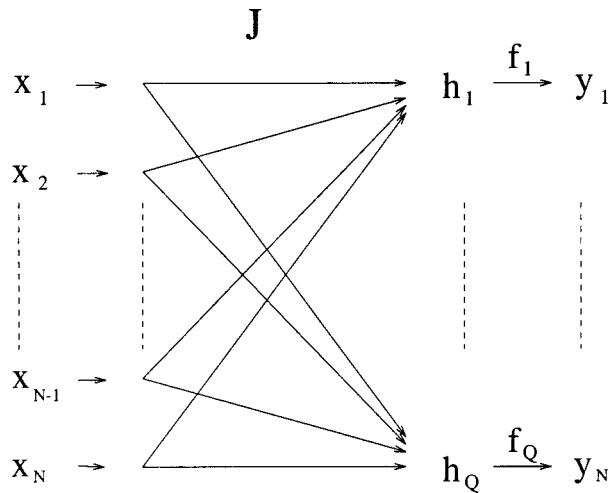


FIG. 2. The component extraction model: the perceptron architecture, where \mathbf{x} is the observation, \mathbf{h} is the extracted component vector, and \mathbf{y} is the network output.

functions is specified in the redundancy quality criterion of (12), the basis functions extracted by ICA, in contrast to PCA, can be nonorthogonal.

A statistical regression model for the extraction model in Eq. (11) has to be specified. For the nonlinear mixture case the regression model needs to be nonlinear in order to simulate G^{-1} . The Multi-Layer Perceptron (MLP), an artificial neural network model, could be chosen for such a case. The nonlinear mixture case will be the subject of future work.

In the present linear mixture case, we use a simpler MLP architecture with no hidden layers as the extraction model. This neural mapping is defined by (from right to left in Fig. 2):

$$\mathbf{y} = f(\mathbf{h}) = f(\mathbf{J} \cdot \mathbf{x}), \quad (13)$$

where f is the logistic function (nonlinear, bounded, and invertible). This model is basically a linear model, but it employs a nonlinear logistic function. A nonlinear f is used, even if the mixture model is linear, for algorithmic considerations (Nadal and Parga 1994, 1997). To obtain statistical independence, the manipulation of higher moments (like $\langle h_i^3 \rangle$, $\langle h_i^4 \rangle$, $\langle h_i^5 \rangle$, ...) is required. Applying nonlinear f_i to the h_i 's allows one to include these higher-order moments because the Taylor expansion of $f(\mathbf{h})$ uses higher powers of the h_i values. So we consider a nonlinear transformation (postfiltering step) on the estimator $\mathbf{h} = \mathbf{J} \cdot \mathbf{x}$: the extracted components are not the output \mathbf{y} or the neural mapping (13), but the vector $\mathbf{h} = \mathbf{J} \cdot \mathbf{x}$.

The parameters, \mathbf{W}_i , defining the matrix \mathbf{J} and the optimal transfer functions f have to be determined by minimizing the redundancy criterion in (12). Practically, it has been demonstrated in various applications (Bell and Sejnowski 1995) that full optimization of the transfer functions is not necessary for performing ICA. Although promising results have been obtained, this anal-

ysis strategy can be improved by introducing some partial adaptation of the transfer functions to the particular problem. We use here the classical sigmoid function $f(x) = 1/(1 + e^{-\beta x})$ that has proven generally useful.

With the information redundancy reduction criterion, (12) and a no-hidden-layer architecture, a straightforward algorithmic implementation of the ICA has been found (Bell and Sejnowski 1995) to estimate the matrix \mathbf{J} :

$$J_{ik}(n+1) - J_{ik}(n) = \rho(n) \left(J_{ik} + \bar{y}_i \sum_l J_{lk} \cdot h_l \right), \quad (14)$$

where $J_{ik}(n)$ is an element of the matrix \mathbf{J} at step n of the gradient descent, ρ is the learning rate parameter of the gradient descent, and

$$\bar{y}_i = \frac{\partial}{\partial y_i} \frac{\partial y_i}{\partial h_i} = \frac{\partial}{\partial h_i} \ln \left(\frac{\partial y_i}{\partial h_i} \right) = 1 - 2y_i. \quad (15)$$

This algorithm is described in a more practical way in the appendix.² Note that, although the theory behind this analysis method may seem complex, the actual computational procedure that results for the linear case is relatively simple.

ICA can be applied to the raw data, $\mathbf{x}(j)$, but it has been shown (Nadal et al. 2000; Aires et al. 2000) that a PCA preprocessing of observations makes the gradient descent step stabler and faster. The N -dimensional data $\mathbf{x}(j)$ are projected onto their first N' (where $N' < N$) principal components using the matrix \mathbf{J}_0 of PCA: observation noise is reduced and fewer free parameters need to be estimated because the dimensions of the matrix \mathbf{J} are considerably reduced, from $Q \times N$ to $Q \times N'$. The dimension of the compression N' is chosen to be equal to Q , the number of extracted components, thus, \mathbf{J} is a $Q \times Q$ matrix. The PCA-compression preprocessing step does not alter the components extracted by ICA if the neglected components (including noise) are statistically weak sources (see Nadal et al. 2000 for a definition of the term "weak"). In this configuration, the ICA can be considered equivalent to performing an oblique rotation of the PCA solution by generalizing decorrelation to statistical independence, but no additional criterion needs to be selected from among a large number of possibilities like the localization constraints in classical RT. The rotation matrix of the PCA solution \mathbf{J}_0 is the ICA solution $Q \times Q$ matrix \mathbf{J} . If the real physical components are Gaussian-distributed, then decorrelation is sufficient and the ICA algorithm would not change the PCA solution since zero gradient is already obtained. In the case where the components are non-Gaussian, the ICA rotates the initial PCA solution. So ICA improves the PCA solution only in the non-Gaussian case. We note that there are many examples

² See also the Computational Neuroscience Laboratory of Terry Sejnowski at The Salk Institute for links to recent literature, software and demos concerning the ICA paradigm: <http://www.cnl.salk.edu/~tewon/ica.cnl.html>.

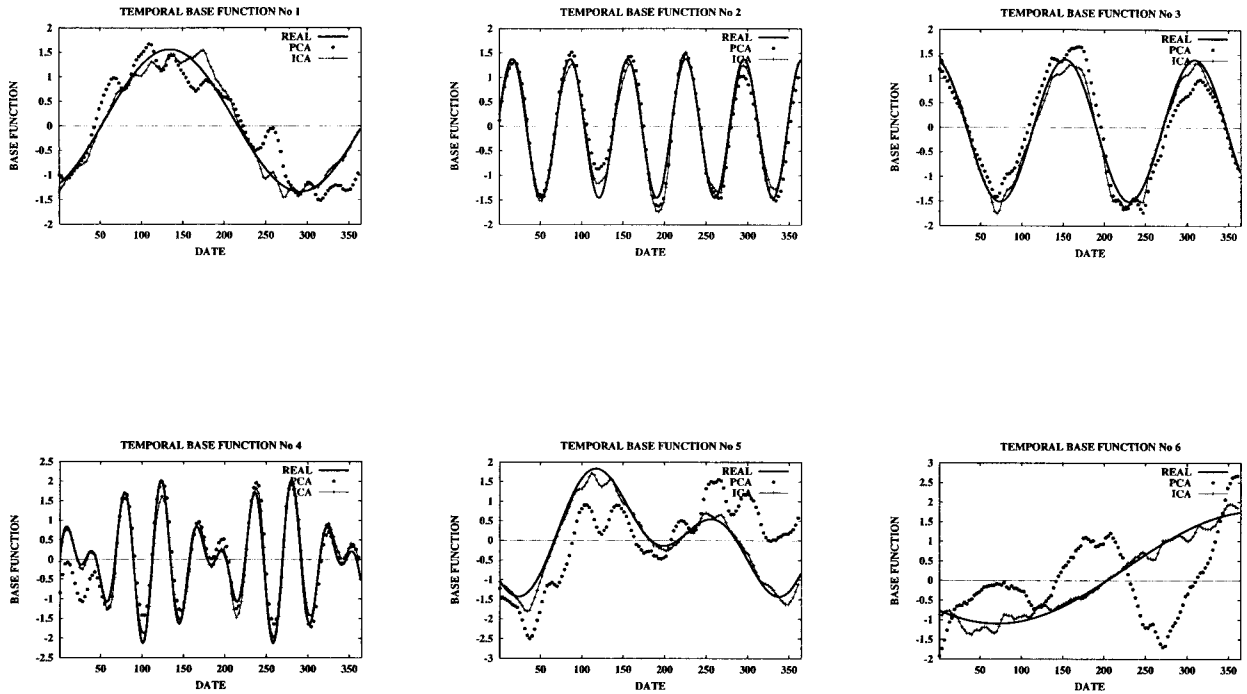


FIG. 3. Temporal basis functions, g_i : ACTUAL (solid lines), PCA estimates (crossed lines), and ICA estimates (dotted lines).

in the literature showing non-Gaussian distributions of climate parameters, for example, cloud optical thickness (Rossow and Schiffer 1991), precipitation water path (Lin and Rossow 1996), and SST (Aires et al. 2000).

4. Application to a linear sum of components

a. Construction of the synthetic dataset

Geophysical time series have been analyzed by linear statistical extraction techniques for decades. The synthetic dataset used in this study is generated to mimic the apparent expectations of such an analysis approach; namely, that the observations are a linear sum of modes with different space and time variations and, so, are separable by such an analysis. Consequently, we exaggerate the space and time differences of some modes, as compared to more realistic atmospheric modes, to make their separation by PCA even easier. On the other hand, we also include two modes that are spatially overlapped, but with different time behaviors, and two spa-

tially separated modes with identical time behavior representing a teleconnection. There are also modes with very different time behavior and two modes are relatively highly correlated on time.

We select $Q = 6$ components representing six different dynamical phenomena, each described by a different temporal basis function, g_i (solid lines in Fig. 3), constructed from composites of sinusoids with different frequencies and phases. Each basis function has been normalized to give a temporal standard deviation of unity. The temporal dimension of these basis functions is taken to be $N = 365$ (e.g., one year of daily data). A spatial resolution of $2.5^\circ \times 2.5^\circ$ is chosen, corresponding to $M = 144 \times 72 = 10\,368$ pixels. Finally, the dataset, $\mathbf{X}_j^i = \{x(j) \in \mathbb{R}^N; j = 1, \dots, M\}$, where \mathbb{R}^N is the space of real vectors of dimension N , $N = 365$ and $M = 10\,368$, is formed from the time series $x(j)$ for each pixel j by the linear sum of the basis functions, $\mathbf{x}(j) = \mathbf{g}_1\sigma_1(j) + \dots + \mathbf{g}_Q\sigma_Q(j) + \varepsilon$ [linear model of Eq. (5)]. The term ε is Gaussian-distributed noise (zero mean and standard-deviation of 0.5), representing very noisy data as might be the case when analysing climate anomalies.

The $\{\sigma_i(j); i = 1, \dots, Q\}$ indicate the strength of each component i at each pixel j , that is, the spatial distributions. These strengths are constructed to have a geographical bell-shaped distribution, giving a different ellipsoidal distribution for each component (left column in Fig. 4). Artificial land contours are introduced into the display of σ_i for easier description of the modes. One of the components has two peaks in its spatial

TABLE 1. Variance explained by Noise, Real, PCA, and ICA components.

Component	Real	PCA	ICA
1	13.3	24.4	12.7
2	12.6	14.5	13.0
3	10.7	10.7	11.3
4	10.7	8.8	10.2
5	10.7	6.2	10.3
6	10.0	3.1	10.0
Noise	33.0	32.3	32.3

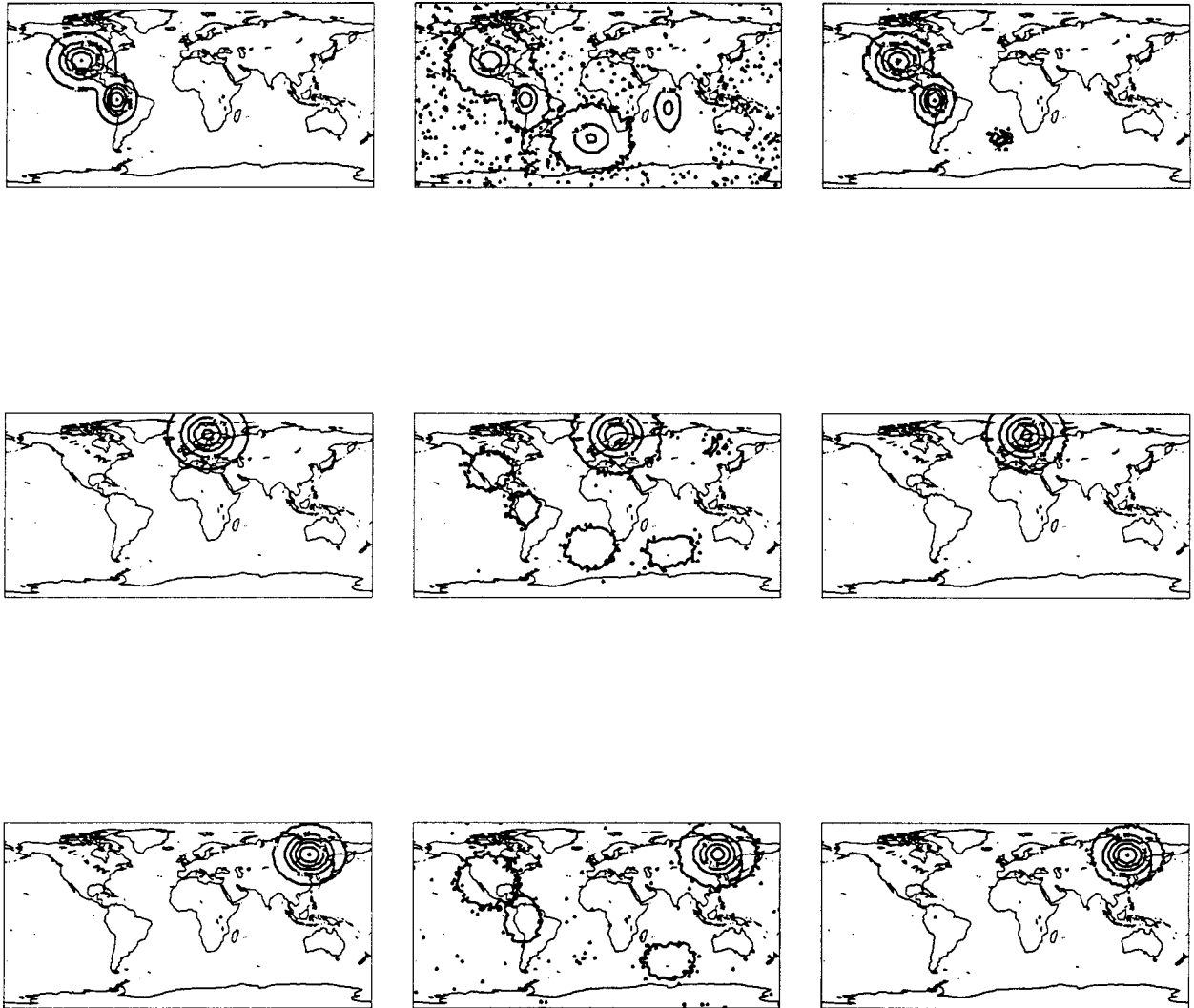


FIG. 4. The maps (left) of the actual components, σ_i , (middle) of the PCA extracted components, h_i , and (right) of the ICA extracted components, h_i ; components number 1–6 from the top to the bottom; component maps have been centered and normalized for comparison purposes. The continental outlines are artificial and used to make discussion of specific features easier.

distribution (near the Americas) to represent a teleconnection pattern (map of component 1 in left column of Fig. 4), so the total number of ellipsoidal peaks is seven. Also, mode 5 is highly correlated in time with mode one, but not perfectly correlated (correlation of the two base functions without noise -0.7). The geographical extent of two of the components overlaps in the Indian Ocean (maps of components 4 and 6 in left column of Fig. 4) to complicate the component extraction process.

The variance contributed by the $Q = 6$ components and the added noise is shown in Table 1: the components produce 67% of the total variance and the noise produces 33%. The total variance of a component results from the combination of the temporal variability of the basis function (as a function of normalized amplitude and frequency) and the spatial extent of the component.

b. Results of PCA and ICA

The PCA components are determined by computing the matrix \mathbf{J}_0 . The best number of PCA components to extract is determined here by observing the spectrum of cumulative percent of variance explained by the PCA components (Fig. 5). More sophisticated criteria have been developed to determine the number of significant components (see, e.g., Jolliffe 1986). The first $Q = 6$ PCA components represent roughly equal variance with a total of 67.7% and the 359 remaining components explain equal portions of the remaining 32.3% of the total variance, representing the noise in the dataset. The PCA temporal basis functions (crossed lines in Fig. 3) are each compared with the real basis function to which it best corresponds. PCA basis functions 2, 3, and 4

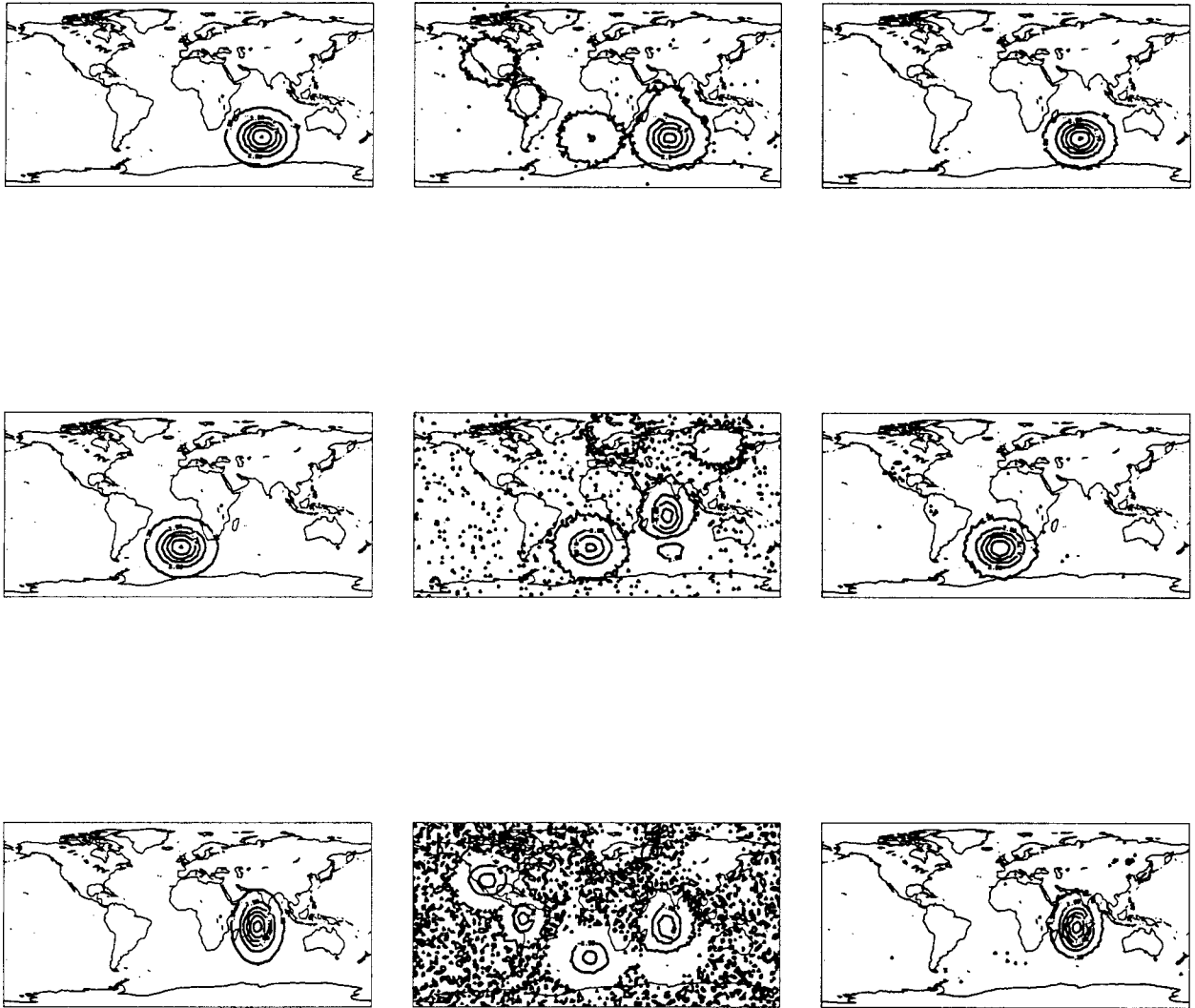


FIG. 4. (Continued)

provide a relatively good estimate of the true functions, although there are some errors near the peak values. PCA temporal basis functions 1, 5, and 6 (low frequencies) are much worse fits. In particular, higher frequencies have been mixed into the real basis functions.

The corresponding PCA component maps are defined at pixel j by the values $(h_1, \dots, h_Q)(j) = \mathbf{J}_0 \cdot \mathbf{X}_j^*$ and are shown in Fig. 4 (middle column), where \mathbf{X}_j^* is the j^{th} column of data matrix \mathbf{X} . We see that the PCA (or EOF) technique confuses elements from the different components, the general mixing problem, such that all of its components exhibit many more geographic peaks than in the real components. Even if the corresponding PCA temporal basis function is relatively well retrieved, the corresponding component map still exhibits the mixing problem (see especially PCA basis function 2 in Fig. 3 and the corresponding PCA component map in Fig. 4). If we had constructed more realistic components—for example, covering the whole Pacific basin, Tropics,

or Northern Hemisphere—the mixing problem would be much less apparent. The fact that PCA mixes these well-separated components means it will mix the less well-separated components of climate variations.

One cause of the mixing is well illustrated in Table 1 where the variance explained by each PCA component is compared to the variance of the actual components. The first PCA component explains 24.4% of the total variance, which is much more than its true variance of 13.3%. The 6th PCA component represents only 3.1%, which is a considerable underestimate of the real value of 10%. Thus, the variance maximization property in PCA shifts signal from other components into the first component, producing a mixture of many true component variabilities. The noise level estimate of 32.3% is a good estimate, but its small underestimate of the real noise is due to the projection of some noise onto the first six PCA components (representing 0.7%).

Particularly notable in Fig. 4 is that the mixing ten-

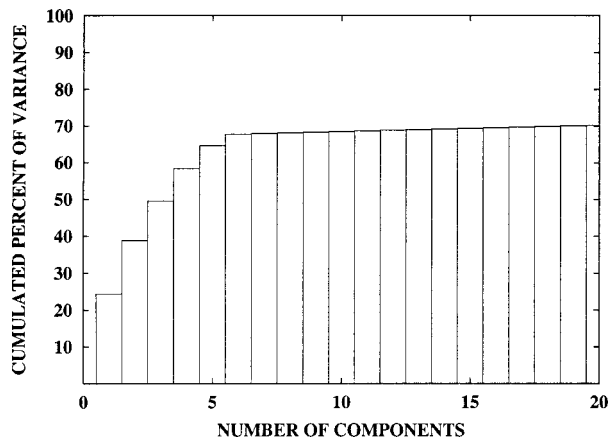


FIG. 5. Cumulative percent of explained variance by the PCA components.

dependency of the (unrotated) PCA could suggest many more teleconnections in observations than are actually present. Since in this synthetic case all six components contribute roughly the same amount of variance (10%–13%), the PCA technique has combined many of the actually separate components into several of its components, trying to maximize the amount of variance explained by each. After the first PCA component, subsequent components often have a mixture of positive and negative values (or loadings) because of the orthogonality constraint. This effect is especially apparent for the overlapping components in the Indian Ocean: two PCA basis functions possess broad central peaks spanning the geographic distribution of both of the real components and two others possess, in this same location, two opposite-signed peaks (see PCA component maps of components 1, 4, 5, and 6 in Fig. 4, middle column). The alternating signs of the poles in the PCA maps, due mainly to the orthogonality constraint, are analogous to the alternating signs of the sinusoid functions in a Fourier analysis. A similar projection of real components into more than one PCA component occurs when a geographically isolated mode moves during the time period (Kim and Wu 1999). Moreover, the component with two peaks near the Americas, representing a real teleconnection, shows up in four of the PCA components (components 1, 3, 4, and 6 in Fig. 4, middle column), but mixed with other components as well, suggesting teleconnections between the Americas and the South Atlantic and Indian Oceans that do not exist. We note also that components 1 and 5, even if they are highly correlated (correlation of 0.7 of the temporal base functions without noise), have been mixed into component PCA 1. Dots in PCA component maps are localized contour lines, showing the sensitivity of PCA component to data noise.

ICA can be applied directly to the raw data $\mathbf{x}(j)$ but, as previously commented in Nadal et al. (2000) and more briefly at the end of section 3, a PCA preprocessing

of observations makes the gradient descent step numerically stabler and faster. So the observed data $\mathbf{x}(j)$ are first projected onto the first $Q = 6$ PCA components using the matrix \mathbf{J}_0 . This has the beneficial effect of removing most of the noise. The ICA technique is then applied to the preprocessed data, $\tilde{\mathbf{x}}(j) = \mathbf{J}_0 \cdot \mathbf{x}(j)$ (dimension $Q = 6$ instead of $N = 365$). As explained in section 3, this is equivalent to performing a rotation on the PCA initial solution where the rotation matrix is the $Q \times Q$ matrix \mathbf{J} of the ICA solution. Thus the six ICA extracted components explain the same amount of variance as the six PCA components (67.7%).

The six ICA basis functions are shown in Fig. 3 (dotted lines). The ICA basis functions are very similar to the real basis functions. This comparison shows how the ICA technique has corrected its first guess (the PCA solution) to be closer to the true solution. The additional information obtained from the requirement for statistical independence is nicely illustrated: the ICA technique has transformed the PCA initial solution for a better retrieval of all six components. The two highly correlated components 1 and 5 have been clearly separated by ICA, illustrating the importance of using the higher order statistics to discriminate such modes. The ICA component maps are presented in Fig. 4 (right column). Generally, the components are well-retrieved and separated, even the teleconnection mode (ICA component 1 in Fig. 4, right column) and the two overlapping modes in the Indian Ocean (ICA components 4 and 6 in Fig. 4, right column). The transformation of the PCA component maps by ICA is always an improvement.

An experiment was conducted with the same data but without the noise. The ICA separates the original six modes almost perfectly and the ICA solution is very close to the real solution. This result indicates that the presence of measurement noise in a dataset will produce a small amount of mode mixing even in the ICA solution; however the results shown here (small hints of other modes in ICA components 4 and 6 in Fig. 4, right column) is produced by a situation where the signal-to-noise ratio is only about two. Although this situation may be relevant to climate studies, ICA can separate most of the noise into its own statistically independent mode.

Table 1 shows that the variance explained by the ICA components is much closer to the real solution than the initial PCA components: the variance explained by the first couple of modes decreases and that retained by the remaining modes increases. Differences between the true and ICA explained variance for each component are less than 0.6%, where the discrepancies are the result of the projection of some part of the noise onto the ICA components.

5. Concluding remarks

For extraction of physically meaningful modes from observations, where the characteristics of the system's

dynamics are not (well-) known, identifying statistically independent variation modes seems to be a sensible alternative for the rotation of a first PCA (i.e., EOF) solution to avoid the PCA mixing problem. Our simple example shows that in the most general, though still linear, case (the most favorable condition for PCA), PCA will mix modes of comparable magnitude, generating spurious regional overlaps or teleconnections where none exists or distorting existing overlaps or teleconnections. In the case of correlated modes, PCA produces mixing by combining the correlated parts of the modes and separating the less correlated parts, spuriously dividing some modes. The PCA is useful in a lot of applications, particularly when the user is interested in only one strong component in the observations that explains the maximum of variance. Note, however, that such use in recent studies requires substantial pre-filtering of the data to isolate such a mode; this is equivalent to application of very strong a priori information about the mode in question. But to extract components when there are several, PCA results will be misleading, even for a simple linear mixture.

We have shown the potential of the ICA technique for separating a complex signal in a more meaningful way. The mixing problem inherent in the PCA technique and the artifacts produced by the orthogonality and maximum-variance constraints of PCA are avoided when rotated by ICA. Moreover, the use of higher-order statistics by ICA to determine statistical independence assumes only the generalization of the decorrelation used in all classical approaches. In some cases, as we have shown, use of higher-order statistics is key to separating similar modes. Nevertheless, even statistical independence does not guarantee that the modes produced by different physical processes will be separated. Like other statistical methods, without a priori information about the actual physical modes in the observations, there is no guarantee that the components extracted have a physical meaning. The user of a particular technique should keep in mind the assumptions used in the technique, the qualities and deficiencies of the method, and be able to put these in the context of each application. The advantage of ICA is its straightforward criterion, that is, the statistical independence of its extracted components.

Some practical disadvantages of ICA exist. As with rotation techniques, it needs an a priori definition of the number of components to extract. But like oblique rotations (Richman 1981), ICA still should be able to extract meaningful components even if more components are extracted than there are actually in the observations (i.e., overfactoring). Practically, the use of higher-order statistics requires many more samples than using second-order statistics, placing stronger demands on the data needed. In addition, the higher-order moments are more sensitive to outliers; however, our use of PCA as a first step in the PCA helps reduce this problem by removing most of the noise and any low frequency-of-occurrence outliers. Computationally, ICA requires

more resources than PCA, but the convergence of the ICA algorithm is very fast.

ICA, by finding statistically independent modes, may provide a better starting point to explore the unknown dynamics of a system. In the case of climate variations, where the components of the system are probably coupled (see, e.g., Salby and Callaghan 2000 or Krishnamurthy and Goswami 2000), considering the modes to be as statistically distinct as possible, even with a linear-ICA, would provide "prototypical" components that might serve as a guide to further investigation. As with the classical PCA technique or classical RT, this first (linear) ICA algorithm is not able to deal correctly with propagating components or components mixed nonlinearly. However, the ICA paradigm (statistical independence) may be a sufficiently powerful concept to be generalized using more advanced statistical models (e.g., more complicated neural networks) to treat nonlinear problems. This requires development of nonlinear solution algorithms and their testing for cases where the combination of modes is nonlinear, when components are physically linked, and for cases with propagating modes.

Acknowledgments. We are grateful to Dr. David Rind and Dr. Ronald L. Miller for their helpful comments. This work was supported by special funding provided by Dr. Robert J. Curran, NASA Climate and Radiation Program.

APPENDIX

Principal Steps of the Algorithm

We adopt here the linear model $\mathbf{x} = \mathbf{G} \cdot \boldsymbol{\sigma}$, where \mathbf{x} is the observation, \mathbf{G} is the basis function matrix, and $\boldsymbol{\sigma}$ is the vector of components to estimate. The goal of the statistical decomposition technique is to estimate a matrix $\mathbf{J} = \mathbf{G}^{-1}$ (the superscript $^{-1}$ represents the pseudoinverse if \mathbf{G} is not square), the filter matrix, using only a dataset of observations $\{\mathbf{x}^e; e = 1, \dots, E\}$, where E is the number of samples in the dataset. With the matrix \mathbf{J} applied to each observation \mathbf{x} , the components $\boldsymbol{\sigma}$ are estimated by $\boldsymbol{\sigma} \approx \mathbf{h} = \mathbf{J} \cdot \mathbf{x}$, and the basis function matrix \mathbf{G} is estimated by the inverse matrix \mathbf{J}^{-1} .

The principal steps of the time series analysis by the ICA technique are the following.

- **Optional preprocessing:** The dataset $\mathbf{X}_t^j = \{\mathbf{x}(j) \in R^N; j = 1, \dots, M\}$, where t is the time index and j is the space index (geographical locations), may require preprocessing: 1) spatial, temporal, or spatio-temporal interpolation to fill in missing data, 2) filtering of data to suppress undesirable frequencies (noise effects), 3) detrending to obtain stationary data, and 4) removing the annual cycle to examine interannual anomalies. None of these steps is required.

- **Chose the space for the decomposition:** 1) Chose

the space in time, which is the approach we have adopted in our study:

$$\mathbf{x}(j) = \mathbf{g}_1\sigma_1(j) + \dots + \mathbf{g}_Q\sigma_Q(j) + \varepsilon;$$

2) in space, 3) in frequency, or 4) in a mixture of these spaces. The observations (a time series or a geographical field, ...) are denoted, in the following, by the d -dimensional vector \mathbf{x}^e and the dataset is $\{\mathbf{x}^e; e = 1, \dots, E\}$.

• **Center the dataset:** The observation mean $\langle \mathbf{x}^e \rangle$ is removed from the dataset: $\mathbf{x}^e \leftarrow \mathbf{x}^e - \langle \mathbf{x}^e \rangle$. This step is necessary for statistical techniques where data are supposed to have zero mean like ICA.

• **Optional normalization:** If the user wants to put the same statistical weight on each coordinate of the observation \mathbf{x}^e ; then the dataset can be normalized by the standard-deviation vector $\mathbf{x}^e \leftarrow \mathbf{x}^e/\mathbf{e}_s$.

• **Optional Eigenvector decomposition:** The covariance (or correlation, in the case of normalized observations) matrix $\langle \mathbf{x}^t \cdot \mathbf{x} \rangle$ is estimated from the dataset. The eigenvalues Λ (diagonal matrix) and the eigenvector matrix \mathbf{V} of $\langle \mathbf{x}^t \cdot \mathbf{x} \rangle$ are then computed using a classical numerical routine. The number of PCA or ICA extracted components Q is chosen by observing the spectrum of eigenvalues.

• **Optional PCA solution:** The PCA solution is computed to preprocess the data.

(i) The $d \times Q$ PCA basis function matrix \mathbf{G}_{PCA} contains in its columns the first Q eigenvectors of \mathbf{V} (the columns of \mathbf{V} represent time series in the time decomposition, and geographical field in the space decomposition, ...).

(ii) Since, by definition, $\mathbf{V}^{-1} = \mathbf{V}^t$, the filter PCA matrix, \mathbf{J}_{PCA} , is equal to the transposed $Q \times d$ basis function matrix, \mathbf{G}_{PCA} . Then, the extracted components \mathbf{h} that estimate the true components $\boldsymbol{\sigma}$ are the projection of the observations \mathbf{x} onto the filters: $\mathbf{h} = \mathbf{J}_{\text{PCA}} \cdot \mathbf{x}$.

(iii) The first Q eigenvalues in Λ represent the variability explained by each of the Q components.

• **ICA solution:**

(i) *Prewhitening of dataset:* The PCA solution is used as a preprocessing step, and the observations \mathbf{x}^e are projected onto the PCA filters $\mathbf{x}^e \leftarrow \mathbf{J}_{\text{PCA}} \cdot \mathbf{x}^e$. The ICA algorithm is then applied into these Q -dimensional data.

(ii) The ICA solution \mathbf{J}_{ICA} is initialized as the identity matrix $\mathbf{I}_{Q \times Q}$. This, associated to the previous whitening step, is equivalent to taking the PCA solution as first guess for ICA.

(iii) For the minimization of the criterion specifying the statistical independence, a gradient descent algorithm is used. The classical gradient descent uses all the samples of the dataset to compute a mean $\Delta J_{ik} = J_{ik}(n+1) - J_{ik}(n)$ in Eq. (14). This algorithm is called the deterministic gradient descent. The major inconvenience of this algorithm is that it can be trapped in local minima. We use, in our application, the stochastic gradient descent algorithm that uses the gradient descent formula (14) iteratively in unique random samples of the dataset.

The stochastic character of the optimization algorithm allows theoretically, and under some constraint not discussed here, for the optimization technique to reach the global minimum of the criterion instead of a local minimum (Duflo 1996).

(iv) An observation \mathbf{x}^e is randomly chosen in the dataset. The propagation through the neural network (chosen model for the component extraction) is given by $\mathbf{y} = f(\mathbf{h}) = f(\mathbf{J}_{\text{ICA}} \cdot \mathbf{x}^e)$, where $f(a) = [1 + \exp(-\beta \cdot a)]^{-1}$ is the logistic function. The parameter β controls the slope of the logistic function, and we take $\beta = 2.0$ as in Bell and Sejnowski (1995). The FORTRAN routine of this process is:

```
c—propagation through the neural network
do i = 1, d
  h(i) = 0.d0
  do k = 1, d
    h(i) = h(i) + J_ICA(i, k)*x^e(k)
  enddo
  h(i) = h(i) + bia(i)
  y(i) = 1.d0/(1.d0 + d exp(-beta*h(i)))
enddo,
```

where *bia* is the classical bias vector in an MLP neural network (not shown in the text for simplicity). We use, in this routine, double precision variables to avoid numerical instabilities.

(v) The learning process is then defined as:

```
c—transitory quantities
do j = 1, d
  hhh(j) = 0.d0
  do k = 1, d
    hhh(j) = hhh(j) + J_ICA(k, j)*h(k)
  enddo
enddo

c—modification of weights
do i = 1, d
  do j = 1, d
    J_ICA(i, j) = J_ICA(i, j) + param*
    & (J_ICA(i, j) + beta*(1.d0 - 2.d0*y(i))*hhh(j))
  enddo
  bia(i) = bia(i) + beta*(1.d0 - 2.d0*y(i))
enddo
```

where *param* is the learning parameter of the gradient descent optimization (we take *param* = 0.0005).

(vi) *Stopping criterion:* Many criteria can be used to define when to stop the learning cycle. The simplest criterion is to determine a priori the number of learning steps. A better criterion is to determine when the difference between solution \mathbf{J}_{ICA} at time t and at time $t+1$ falls below some threshold value. Another stopping criterion is to evaluate the statistical independence of the extracted components. Here, \mathbf{h} cumulants (i.e., additive higher-order moments) are a practical way to do that, but this approach is computationally expensive.

The learning algorithm returns to step 4 until the stopping criterion is reached.

• **Analysis of results:** When the matrix \mathbf{J}_{ICA} has been determined by ICA, the global ICA filters (taking into account the PCA preprocessing) are defined by the $Q \times d$ matrix: $\mathbf{J}_{GLO} = \mathbf{J}_{ICA} \cdot \mathbf{J}_{PCA}$.

(i) The projection of data is used to estimate the components: $\mathbf{h} = \mathbf{J}_{GLO} \cdot \mathbf{x}^e$.

(ii) The $d \times Q$ ICA basis function matrix $\mathbf{G}_{GLO} = \mathbf{J}_{GLO}^{-1} = \mathbf{G}_{PCA} \cdot \mathbf{J}_{ICA}^{-1}$ is normalized to obtain normalized ICA basis functions, as in PCA approach.

(iii) Computation of explained variance of each of the basis functions.

REFERENCES

- Aires, F., A. Chédin, and J.-P. Nadal, 2000: Independent component analysis of multivariate times series: Application to the tropical SST variability. *J. Geophys. Res.*, **105** (D13), 17 437–17 455.
- Atick, J. J., 1992: Could information theory provide an ecological theory of sensory processing? *Network Comput. Neural Syst.*, **3**, 213–251.
- Bell, A. J., and T. J. Sejnowski, 1995: An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.*, **7**, 1004–1034.
- Broomhead, D., and G. King, 1986: Extracting qualitative dynamics from experimental data. *Physica D*, **20**, 217–236.
- Buell, C. E., 1975: The topography of empirical orthogonal functions. Preprints, *Fourth Conf. on Probability and Statistics in Atmospheric Science*, Tallahassee, FL, Amer. Meteor. Soc., 188–193.
- Burel, G., 1992: Blind source separation of sources: A nonlinear algorithm. *Neural Networks*, **5**, 937–947.
- Burgers, G., and D. B. Stephenson, 1999: The “Normality” of El Niño. *Geophys. Res. Lett.*, **26**, 1027–1030.
- Cardoso, J. F., C. Jutten, and P. Loubaton, Eds., 1999: Preprints, *First Int. Workshop on Independent Component Analysis and Signal Separation*, Aussois, France, ICA, 506 pp.
- Chung, C., and S. Nugam, 1999: Weighting of geophysical data in Principal Component Analysis. *J. Geophys. Res.*, **104**, 16 925–16 928.
- Comon, P., 1994: Independent Component Analysis: A new concept? *Signal Process.*, **36**, 287–314.
- Dacunha-Castelle, D., and M. Duffo, 1982: Probabilités et Statistiques, Tome 1: Problèmes à Temps Fixe. Masson, France.
- Duffo, M., 1996: *Algorithmes Stochastiques: Mathématiques et Applications*. Springer Verlag, 319 pp.
- Horel, J., 1981: A rotated principal component analysis of the interannual variability of the northern hemisphere 500-mb height field. *Mon. Wea. Rev.*, **109**, 2080–2092.
- Hyvärinen, A., and E. Oja, 2000: Independent component analysis: Algorithms and applications. *Neural Networks*, **13**, 411–430.
- Jolliffe, I. T., 1986: *Principal Component Analysis*. Springer Verlag, 271 pp.
- Jutten, C., and J. Herault, 1991: Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture. *Signal Process.*, **24**, 1–10.
- Karl, T. R., A. J. Koscielny, and H. F. Diaz, 1982: Potential errors in the application of principal component (eigenvector) analysis to geophysical data. *J. Appl. Meteor.*, **21**, 1183–1186.
- Kim, K.-Y., and Q. Wu, 1999: A comparison study of EOF techniques: Analysis of nonstationary data with periodic statistics. *J. Climate*, **12**, 185–199.
- Korres, G., N. Pinardi, and A. Lascaratos, 2000: The ocean response to low-frequency interannual atmospheric variability in the Mediterranean Sea. Part II: Empirical orthogonal functions analysis. *J. Climate*, **13**, 732–745.
- Krishnamurthy, V., and B. N. Goswami, 2000: Indian Monsoon–ENSO relationship on interdecadal timescale. *J. Climate*, **13**, 579–595.
- Lin, B., and W. B. Rossow, 1996: Seasonal variation of liquid and ice water path in nonprecipitating clouds over oceans. *J. Climate*, **9**, 2890–2902.
- Lorenz, E., 1951: Seasonal and irregular variations of the northern hemisphere sea-level pressure profile. *J. Meteor.*, **8**, 52–59.
- Monahan, A. H., 2000: Nonlinear principal component analysis by neural networks: Theory and application to the Lorenz System. *J. Climate*, **13**, 821–835.
- Nadal, J.-P., and N. Parga, 1994: Nonlinear neurons in the low-noise limit: A factorial code maximizes information transfer. *Comput. Neural Syst.*, **5**, 565–581.
- , and —, 1997: Redundancy reduction and independent component analysis: Conditions on cumulants and adaptive approaches. *Neural Comp.*, **9**, 1421–1456.
- , E. Korutcheva, and F. Aires, 2000: Blind source separation in the presence of weak sources. *Neural Networks*, **13**, 589–596.
- Palmer, T. N., 1999: A nonlinear dynamical perspective on climate prediction. *J. Climate*, **12**, 575–591.
- Press, W. H. P., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in Fortran*. Cambridge University Press, 973 pp.
- Richman, M., 1981: Obliquely rotated principal components: An improved meteorological map typing technique? *J. Appl. Meteor.*, **20**, 1145–1159.
- , 1986: Rotation of principal components. *J. Climatol.*, **6**, 293–335.
- Rossow, W. B., and R. A. Schiffer, 1991: ISCCP cloud data products. *Bull. Amer. Meteor. Soc.*, **72**, 2–20.
- Salby, M., and P. Callaghan, 2000: Connection between the solar cycle and the QBO: The missing link. *J. Climate*, **13**, 328–338.
- Vautard, R., P. Yiou, and M. Ghil, 1992: Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. *Physica D*, **58**, 95–126.
- , C. Pires, and G. Plaut, 1996: Long-range atmospheric predictability using space–time principal components. *Mon. Wea. Rev.*, **124**, 288–307.
- von Storch, H., and C. Frankignoul, 1998: Empirical modal decomposition in coastal oceanography. K. H. Brink and A. R. Robinson, Eds., *The Sea*, Vol. 16, Wiley and Sons, 419–455.
- Xue, Y., A. Leetmaa, and M. Ji, 2000: ENSO Prediction with Markov Models: The impact of sea level. *J. Climate*, **13**, 849–871.