Modeling multivariable hydrological series: Principal component analysis or independent component analysis?

Seth Westra,[1] Casey Brown,[2] Upmanu Lall,[2] and Ashish Sharma[1]

Received 11 October 2006; revised 6 February 2007; accepted 20 February 2007; published 29 June 2007.

[1] The generation of synthetic multivariate rainfall and/or streamflow time series that accurately simulate both the spatial and temporal dependence of the original multivariate series remains a challenging problem in hydrology and frequently requires either the estimation of a large number of model parameters or significant simplifying assumptions on the model structure. As an alternative, we propose a relatively parsimonious two-step approach to generating synthetic multivariate time series at monthly or longer timescales, by first transforming the data to a set of statistically independent univariate time series and then applying a univariate time series model to the transformed data. The transformation is achieved through a technique known as independent component analysis (ICA), which uses an approximation of mutual information to maximize the independence between the transformed series. We compare this with principal component analysis (PCA), which merely removes the covariance (or spatial correlation) of the multivariate time series, without necessarily ensuring complete independence. Both methods are tested using a monthly multivariate data set of reservoir inflows from Colombia. We show that the discrepancy between the synthetically generated data and the original data, measured as the mean integrated squared bias, is reduced by 25% when using ICA compared with PCA for the full joint distribution and by 28% when considering marginal densities in isolation. These results suggest that there may be significant benefits to maximizing statistical independence, rather than merely removing correlation, when developing models for the synthetic generation of multivariate time series.

Citation: Westra, S., C. Brown, U. Lall, and A. Sharma (2007), Modeling multivariable hydrological series: Principal component analysis or independent component analysis?, *Water Resour. Res.*, *43*, W06429, doi:10.1029/2006WR005617.

1. Introduction

[2] An important objective in stochastic hydrology is to generate synthetic rainfall and/or streamflow sequences that have similar statistics and dependence structures to those of the historical record. These sequences represent plausible future rainfall and/or streamflow scenarios which can be used as inputs in a range of applications, such as the design and operation of reservoirs, irrigation systems and hydroelectric systems.

[3] A large volume of literature exists on modeling single variable (univariate) hydrological time series, of which the autoregressive (AR) and autoregressive moving average (ARMA) class of models are arguably the most common [*Box et al.*, 1994; *Bras and Rodrigues-Iturbe*, 1985; *Loucks et al.*, 1981; *Salas*, 1992], particularly for time series of monthly or greater timescales. These are parametric models that seek to preserve the mean, standard deviation and correlation structure of the original time series, under the assumption that the data are normally distributed, which frequently necessitates that the data be transformed prior to analysis. Alternatively, a number of nonparametric approaches are available [e.g., *Lall et al.*, 1996; *Lall and Sharma*, 1996; *Sharma*, 2000; *Sharma and O'Neill*, 2002; *Sharma et al.*, 1997] which do not require prior assumptions on the nature of the probability distribution.

[4] The situation becomes more complicated when considering the multivariate case. This is because, in addition to simulating temporal dependence, it is also necessary to focus on maintaining spatial dependence. A number of multivariate methods exist, such as a multivariate extension to the ARMA suite of models, which seek to maintain the covariance structure of the observed time series [*Pegram and James*, 1972; *Salas*, 1992; *Wilks*, 1995]. The trouble with such approaches is that it is usually necessary to estimate a large number of parameters, which can render the approach considerably more difficult to apply compared to univariate methods. To simplify parameter estimation, it is sometimes possible to diagonalize the covariance matrix using a technique such as principal components analysis (PCA) to decouple the multivariate time series into component univariate models so that model parameters do not have to be estimated jointly [*Salas*, 1992]. A more fundamental difficulty with the multivariate ARMA model is that, as parameters are often estimated using the method of

[1] School of Civil and Environmental Engineering, University of New South Wales, Sydney, New South Wales, Australia.
[2] Department of Earth and Environmental Engineering, Columbia University, New York, New York, USA.

0043-1397/07/2006WR005617$09.00

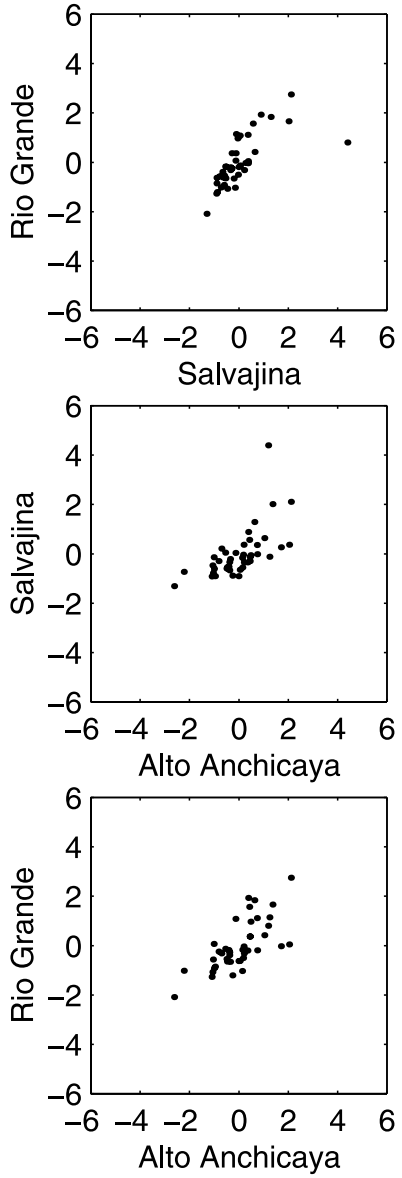

Figure 1. Example of a trivariate streamflow data set located in Colombia, plotted as bivariate pairs. The data have been normalized and highlight a number of issues common to the synthetic generation of a multivariate data set, including sparsity of the data set (42 data points), non-Gaussianity of the joint and marginal probability densities, and possible nonlinear dependence between the data.

moments, involving maintaining order one and two moment statistics, it only considers the covariance structure of the multivariate time series. Although the covariance is an important measure of dependence, it in general does not capture the complete dependence structure of the stochastic process [*Waymire and Gupta*, 1981].

[5] As an example of some of the difficulties with using traditional multivariate methods, we present December streamflow time series in three locations in Colombia in Figure 1, which highlight some common issues with capturing joint dependence. Some notable challenges using traditional multivariate methods for this type of data include: (1) the sparsity of the data, with each time series containing only 42 data points, making parameter estimation for the multivariate case difficult; (2) the possible non-Gaussian distribution of the individual time series, which would necessitate the additional step of finding a suitable transformation to the data; and (3) the possible nonlinear dependence between the individual time series.

[6] To model a data set of this nature, it is desirable to have a method which is parsimonious, and does not make assumptions on the probability distribution of individual variables or the nature of the spatial dependence among them. In this paper we present such an approach in which a rotation is applied to the multivariate data set with the aim of minimizing an estimate of dependence (mutual information) [see *Fraser and Swinney*, 1986] between the rotated series. This technique, known as independent component analysis, is related to the PCA approach described above, except that whereas PCA seeks only to diagonalize the covariance matrix, ICA is also capable of minimizing higher-order dependence. Once a rotation is found that minimizes the dependence of the rotated components, it is possible to consider each of the time series as a univariate case, so that only the temporal characteristics of the time series need to be considered. The inverse of the original rotation is then applied to the synthetically generated univariate series, to ensure that the spatial dependence is preserved.

[7] The remainder of this paper is organized as follows. In section 2, we provide an overview of the mathematical basis of both PCA and ICA, and illustrate the importance of considering not only covariance, but also higher-order moments, when seeking a statistically independent representation of the data set. Section 3 provides an overview of the hydrologic data set used in the analysis. The benefits of considering higher-order statistics are highlighted in section 4 after applying both PCA and ICA to the hydrologic data.

2. Component Extraction Techniques

2.1. Principal Component Analysis

[8] Principal component analysis (PCA) is a widely used component extraction technique that focuses on providing a representation of a multivariate data set using the information that is contained within the covariance matrix, so that the extracted components are mutually uncorrelated. In addition, the principal components have the important property that successive components explain the maximum residual variance of the data in a least squares sense. For these reasons, an important application of the PCA technique is to reduce the dimension of the original data set, by retaining only those principal components that explain a significant portion of the data variance.

[9] To explain the PCA method, we first define $\mathbf{x}$ as the m-dimensional column vector of observations that have been centered so that $\mathbf{x} = \mathbf{x_0} - E\{\mathbf{x_0}\}$, where $\mathbf{x_0}$ represents the original, noncentered data set. The solution to the PCA problem is then simply defined in terms of the unit-norm eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_m$ of the covariance matrix $\mathbf{C_x} = E\{\mathbf{x}\mathbf{x}^\mathbf{T}\}$, which have been ordered so that the corresponding eigenvalues $d_1, \ldots, d_m$ satisfy $d_1 \geq d_2 \geq \ldots \geq d_m$. The first principal component of $\mathbf{x}$ may now be written as

$$\mathbf{PC}_1 = \mathbf{e}_1^\mathrm{T}\mathbf{x} \tag{1}$$

with successive PCs defined in a similar fashion. The solution to the PCA problem therefore requires only the use of classic algebraic methods to find the eigenvectors and corresponding eigenvalues of $\mathbf{x}$. (For computational aspects of eigenvector/eigenvalue estimation, refer to *Golub and Van Loan* [1996].) Because of the ordering of the eigenvectors and eigenvalues, reducing the dimension of the data set to dimension n, with $n \leq m$, is now trivial and simply involves discarding all principal components of order greater than n.

[10] A related method is known as whitening, which not only requires that the components are mutually uncorrelated, but also that the variances of the extracted components are equated to unity. Letting $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_n)$ be the matrix whose columns are the unit norm eigenvectors, and $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_n)$ be the diagonal matrix of the eigenvalues of $\mathbf{C_x}$, then the linear whitening transform is given by

$$\mathbf{V} = \mathbf{D}^{-1/2}\mathbf{E}^{\mathrm{T}} \tag{2}$$

This matrix can always be found, and as will be discussed in the following section, is an important preprocessing step for independent component analysis [*Hyvarinen et al.*, 2001].

2.2. Overview of ICA

[11] The ICA method, first introduced by *Herault and Jutten* [1986], may be considered as an extension to PCA [*Oja*, 2004], except that whereas PCA focuses on identifying components based only on second-order statistics (covariance), ICA also considers higher-order properties which allows it to search for components that are statistically mutually independent. ICA already has been applied successfully in a wide range of areas, including blind source separation (BSS) and feature extraction [see *Hyvarinen*, 1999, and references therein; *Lee*, 1998].

[12] The simplest form of ICA occurs when an m-dimensional observation vector, $\mathbf{x} = (x_1, \ldots, x_m)^{\mathrm{T}}$ of length l is derived through the mixing of an n-dimensional "source" vector, $\mathbf{s} = (s_1, \ldots, s_n)^{\mathrm{T}}$, also of length l, commonly referred to as the independent components [*Comon*, 1994]. These ICs are assumed to be non-Gaussian (with the possible exception of at most one IC, since by knowing all but one IC, the final IC can be specified automatically), mutually statistically independent and zero mean. In addition, it is assumed that $n \leq m$. Put into vector-matrix notation, and assuming that the mixing is both linear and stationary, yields

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{3}$$

where $\mathbf{A}$ is known as the mixing matrix of dimension $m \times n$. The objective of ICA is to estimate the mixing matrix, $\mathbf{A}$, as well as the independent components, $\mathbf{s}$, knowing only the observations $\mathbf{x}$. This can be achieved up to some scalar multiple of $\mathbf{s}$, since any constant multiplying an independent component in equation (3) can be cancelled by dividing the corresponding column of the mixing matrix $\mathbf{A}$ by the same constant.

[13] Central to the identification of the ICs from the data $\mathbf{x}$ is the assumption that all except at most one IC will be "maximally non-Gaussian" [*Hyvarinen et al.*, 2001]. This follows from the logic outlined in the central limit theorem, which is that if one mixes independent random variables through a linear transformation, the result will be a set of variables that tend to be Gaussian. If one reverses this logic, it can be presumed that the original independent components must have a distribution that has minimal similarity to a Gaussian distribution. Consequently, the approach adopted to extract ICs from data containing mixed signals amounts to finding a transformation that results in variables that exhibit maximal non-Gaussianity as defined through an appropriately specified statistic.

[14] The principal advantage of ICA over PCA is that ICA results in components that are independent, whereas PCA leads to components that, while being uncorrelated, may exhibit strong dependence on each other. The independent components are extracted using higher-order information, i.e., information other than that contained in the covariance matrix of $\mathbf{x}$ [*Oja*, 2004]. PCA remains a valuable preprocessing step, however, both as a means of dimension reduction, and as a starting point for whitening (or sphering) the data such that $\mathbf{x}$ is linearly transformed into another n-dimensional vector $\mathbf{z}$ that has a unit covariance matrix. Although the additional step of minimizing dependence increases the computational load involved in estimating components, the increase in computing time was found to be relatively insignificant for the short data sets considered in this paper.

2.3. Example

[15] To illustrate some important differences between PCA and ICA, and demonstrate why it is necessary also to account for higher-order dependence when separating signals for synthetic time series generation, we present the following example. The example is set up by combining two independent signals each of length 5000 which each have a uniform distribution $U[-\sqrt{3}, \sqrt{3}]$, denoted by s_1 and s_2. The bounds of the uniform distribution were selected so that the signals have unit variance. The mixed components, x_1 and x_2, are derived through equation (3) using the following mixing matrix:

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix} \tag{4}$$

These mixed components exhibit joint dependence and are what we refer to as the "observed" data set, analogous to the multivariate streamflow data that we will be considering later in the paper. The joint probability density function of x_1 and x_2 is shown in Figure 2a. As can be seen, the mixed signals appear to be much closer to a normal distribution compared to the original signals, which by construction are uniformly distributed. This is anticipated because of the central limit theorem, which states that under certain conditions, the distribution of a sum of independent random variables tends toward a Gaussian distribution [*Hyvarinen et al.*, 2001]. The eigenvectors are also shown, with $\mathbf{e_1}$ representing the direction of maximum variance, and $\mathbf{e_2}$ constrained to be orthogonal to $\mathbf{e_1}$.

[16] We now use a linear transform, $\mathbf{V}$, to whiten the data such that the elements of $\mathbf{z}$ are mutually uncorrelated, and all have unit variance. We thus have the following relationships between the independent components, $\mathbf{s}$, the observed variables, $\mathbf{x}$, and the whitened data, $\mathbf{z}$:

$$\mathbf{z} = \mathbf{V}\mathbf{x} = \mathbf{V}\mathbf{A}\mathbf{s} \tag{5}$$

The whitened variables are shown in Figure 2b. The marginal probability distributions are clearly not uniformly distributed, and demonstrate that the original source signals, **s**, still have not been found. The advantage of this process, however, is that since we are now dealing with a transformed variable whose elements have zero mean, are mutually uncorrelated and have unit variance, the ICA solution is limited to some orthogonal rotation of the whitened data set about the origin. Denoting a unit vector defining a line passing through the origin of the data in Figure 2b by **w**, then the projection of **z** on the line is given by $\mathbf{y} = \mathbf{w}^T\mathbf{z}$. It has been shown [*Oja*, 2004] that because of prewhitening the data, no matter what the angle of the projection, it always holds that **y** has zero mean and unit variance.

[17] The object of ICA therefore is to find a suitable vector, **w**, that ensures the resulting components **y** are independent, which is obtained through a maximization of the higher-order moments of $\mathbf{w}^T\mathbf{z}$ as described in section 2.4. The solution is shown in Figure 2c, and illustrates that the optimum solution recovers the uniform distribution of the original signals. This process is repeated until all the ICs are found, and **w** approximates one of the rows of the matrix $\mathbf{[VA]}^{-1}$. Thus **y** becomes an estimator of the original independent components, **s**.

2.4. Estimating the Independent Components

[18] As mentioned previously, the objective of ICA is to find projections which yield components that are as independent as possible, where independence means that the joint probability density function can be factorized as follows:

$$f(y_1, y_2, ..., y_n) = f_1(y_1) f_2(y_2) ... f_n(y_n) \quad (6)$$

where f_i represents the marginal probability density function of y_i, and $f(y_1, y_2, \ldots, y_n)$ represents the joint probability density function of all the y_i. It has furthermore been illustrated that, as inferred from the central limit theorem, this objective is equivalent to finding the directions of maximum non-Gaussianity. ICA can thus be thought of as consisting of two basic elements:

[19] 1. Identifying some measure of the non-Gaussianity of the projection $\mathbf{w}^T\mathbf{z}$, often referred to as an objective function or contrast function.

[20] 2. Finding some algorithm that will optimize this non-Gaussianity.

[21] One of the most commonly used measures of non-Gaussianity is kurtosis, however this measure has been found to be highly sensitive to outliers [*Hyvarinen*, 1997], and hence has not been used in this study. An alternative method uses a quantity known as negentropy [*Comon*,

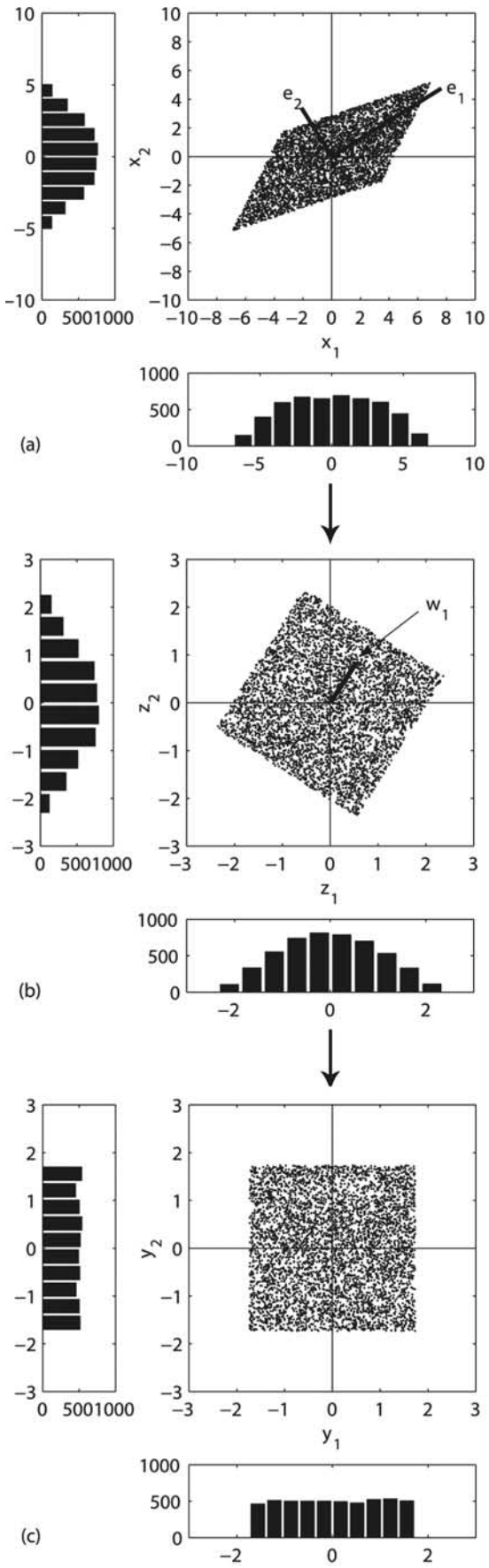

Figure 2. (a) Plot of $\mathbf{x} = [x_1, x]$. The directions of eigenvectors $\mathbf{e_1}$ and $\mathbf{e_2}$ are shown and represent the principal directions of the bivariate data set. The principal components are the projections of **x** onto the principal directions $\mathbf{e_1}$ and $\mathbf{e_2}$. (b) Plot of the whitened data time series $\mathbf{z} = [z_1, z_2]$. To find the ICA solution, we search for a vector **w** such that the projection $\mathbf{y} = \mathbf{w}^T\mathbf{z}$ has maximum non-Gaussianity. (c) Plot of the estimated independent components, $\mathbf{y} = [y_1, y_2]$. The probability distributions are also shown and approximate a normal distribution.

1994; *Hyvarinen et al.*, 2001], and is based on the information theoretic result that a Gaussian variable has the largest entropy of all random variables of equal variance. Negentropy J for a random variable $\mathbf{y}$ is defined as

$$J(\mathbf{y}) = H(\mathbf{y}_{\text{gauss}}) - H(\mathbf{y}), \tag{7}$$

where $H(\mathbf{y})$ is the differential entropy of $\mathbf{y}$ defined as

$$H(\mathbf{y}) = -\int f(\mathbf{y}) \log f(\mathbf{y}) d\mathbf{y} \tag{8}$$

and $H(\mathbf{y}_{\text{gauss}})$ is similarly defined as the entropy of a Gaussian variable $\mathbf{y}_{\text{gauss}}$ of the same covariance matrix as $\mathbf{y}$ [*Comon*, 1994]. The benefit of this quantity is that it is always nonnegative, zero only when $\mathbf{y}$ has a Gaussian distribution, and that estimation of the ICs by maximization of this quantity is equivalent to the minimization of the mutual information, provided that the estimates are constrained to be uncorrelated [*Hyvarinen et al.*, 2001]. The problem, however, is that using negentropy is computationally difficult, as it requires the estimate of the probability density function of $\mathbf{y}$ as shown in equation (8). An approximation of negentropy can be used instead, which is given as

$$J(\mathbf{y}) \approx \left[E\{G(\mathbf{y})\} - E\left\{G\left(\mathbf{y}_{\text{gauss}}\right)\right\}\right]^2 \tag{9}$$

where G is an appropriately chosen nonlinear function often referred to as a contrast function, and the second term in the parentheses is a normalization constant that makes the negentropy J equal to zero if $\mathbf{y}$ has a Gaussian distribution. It has been shown that G can be almost any nonquadratic, well behaving even function [*Hyvarinen and Oja*, 1998]. In the present case, we use

$$G(\mathbf{y}) = \log\cosh(\mathbf{y}) \tag{10}$$

as this is regarded as a good general purpose contrast function because of its convergence properties and robustness against outliers [*Hyvarinen*, 1997]. Since the second term in equation (9) is constant, maximizing non-Gaussianity for a projection $y = \mathbf{w}^T\mathbf{z}$ can be achieved simply by looking at the extrema of the contrast function $E\{G(y)\} = E\{G(\mathbf{w}^T\mathbf{z})\}$ over the unit sphere $||\mathbf{w}||$. The FastICA algorithm [*Hyvarinen and Oja*, 1997] has been found to be an efficient method for finding this extrema, with the central updating rule given as

$$\mathbf{w} \longleftarrow E\left\{G'\left(\mathbf{w}^T\mathbf{z}\right)\mathbf{z}\right\} - E\left\{G''\left(\mathbf{w}^T\mathbf{z}\right)\right\}\mathbf{w} \tag{11}$$

where G′ and G″ are the first and second derivatives of G, respectively. The iteration of this updating rule constrained to the unit sphere is sufficient for estimating one of the ICs. To estimate all the independent components, it is necessary to orthogonalize the vectors $\mathbf{w}$ after each iteration. Here we use a method known as symmetric orthogonalization, which estimates each of the individual vectors w_i, with $i = 1, \ldots, n$, in parallel. This ensures that estimation errors in finding the first vector are not cumulated in subsequent vectors, with further details provided by *Hyvarinen et al.* [2001].

[22] The algorithm that we use for the present analysis can now be summarized as follows [*Hyvarinen et al.*, 2001]:

[23] 1. Center the data set: $\mathbf{x} = \mathbf{x_0} - E\{\mathbf{x_0}\}$.

[24] 2. Prewhiten the data set using $\mathbf{z} = \mathbf{V}\mathbf{x}$.

[25] 3. Choose n, the number of independent components to estimate.

[26] 4. Choose random initial values for the $\mathbf{w}_i$, $i = 1, \ldots, n$, each of unit norm.

[27] 5. Do a symmetric orthogonalization of the matrix $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)^T$ by $\mathbf{W} <- (\mathbf{W}\mathbf{W}^T)^{-1/2}\mathbf{W}$.

[28] 6. For every $i = 1, \ldots, n$, estimate $\mathbf{w}_i$ using equation (11).

[29] 7. If not converged, go back to step 5.

2.5. Using ICA for Stochastic Generation

[30] To illustrate the utility of ICA in the generation of synthetic rainfall and/or streamflow time series, and to highlight the importance of higher-order statistics in finding independent univariate representations of the data set, we present the following example. We commence with the PCA and ICA solutions to the same two-dimensional uniform data set that was presented in section 2.3, shown again as blue dots in Figures 3a and 3b, respectively. The marginal distributions are shown as histograms, and represent the probability density estimates of the transformed, univariate time series.

[31] The synthetically generated data are shown as red dots in Figures 3a and 3b, and are obtained by applying a univariate bootstrap with replacement to each of the transformed time series. The bootstrap was selected to show the effect of considering the transformed univariate time series independently from each other in a synthetic modeling framework, however a range of univariate time series approaches that capture temporal dependence would be expected to yield similar results, provided that the marginal probability density estimates are maintained.

[32] The results of this analysis demonstrate that, as expected, the joint density is captured only for the ICA solution, whereas the PCA solution results in a significantly different joint density to the original data, since this solution is not truly statistically independent. More interesting is the effect when the bootstrapped data are rotated back to the original data space via the inverse PCA or ICA transforms, shown as red dots in Figures 3c and 3d, respectively. Once again, only in the case of ICA is the joint density maintained. In the case of the PCA solution, however, not only is the joint density represented inaccurately, but the marginal density estimates also do not reflect the true marginal density, with the tails of the synthetically generated data set being markedly longer than for the original data. If the above conclusions also can be applied to the synthetic generation of rainfall and/or streamflow data, then the consideration of only second-order dependence could result in significant distortions in both the spatial dependence and the representation of extreme events.

[33] The example above presents an informal argument for using ICA as the basis for decomposing the multivariate data for synthetic generation. A more rigorous approach is described in section 4 of this paper, using univariate and multivariate kernel density estimates to compare the marginal and joint density of the original time series with the marginal and joint density of the ICA and PCA-derived

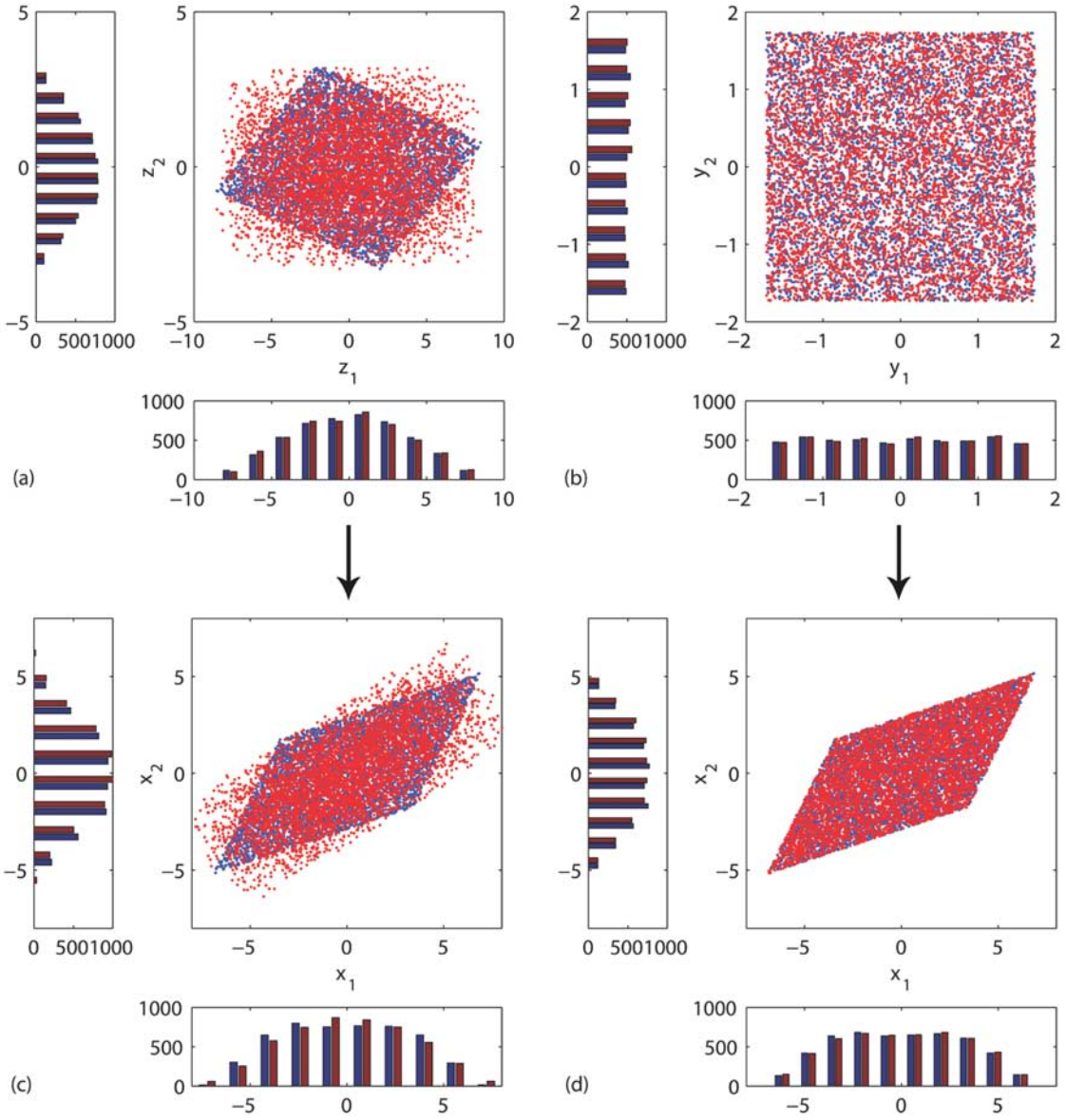

Figure 3. Illustration of the importance of higher-order statistics in synthetically generating multivariate data: (a) PCA-transformed and (b) ICA-transformed data (blue dots), with data generated synthetically by bootstrapping from the marginal distributions shown as red dots. As expected, even though for both cases the marginal distribution is approximately maintained, only the solution from ICA is able to accurately capture the joint dependence. The inverse transformation is then applied to both the (c) PCA and (d) ICA solutions. Here, for the PCA solution, neither the joint nor the marginal distributions are accurately simulated, showing that focusing only on correlation statistics can have profound implications in terms of incorrectly reflecting the joint and marginal statistics from the original data set.

synthetically generated data, respectively. A multivariate streamflow data set of reservoir inflows from Colombia is used for this analysis, and this data set is described in more detail in the following section.

3. Data

[34] Streamflow data at 20 stations in Colombia have been used for this analysis, the locations of which are shown in Figure 4. Of these streamflow stations, a more detailed analysis of three rivers: the Rio Grande, the Cuaca at the Salvajina station and the Alto Anchicaya, was conducted to illustrate the method discussed in this paper. Summary statistics for these three stations are provided in Table 1. Each data station is located at a reservoir inflow point and represents observed inflows to three hydroelectric generating reservoirs. Measurements were initiated during the design phase of each proposed reservoir/dam and continue to the present. Each station retains the name of the river on which it is located. The stations are located in one of the three major basins that drain the country, the Magdalena-Cauca, which flows north to the Caribbean Sea. The other two are the Orinoco basin, which also flows to the Caribbean and the Amazon, and are not represented by this data

Figure 4. Streamflow stations used in the analysis.

set. The data represent unimpaired flows, which in certain cases (e.g., the Alto Anchicaya from 1963–1975) has been calibrated using regions with longer and more reliable records.

[35] The critical hydrologic season for hydroelectric generation is October – November – December. During January – February – March reservoir inflow is low and the system relies on storage of inflows that occur in the previous months. In this study, we use December streamflow as representative of the critical reservoir inflow period.

4. Method and Results

[36] The aim of this paper is to demonstrate how ICA can be used to reduce a multivariate time series into a set of statistically independent univariate time series, so that spatial dependence and temporal dependence can be considered in separate models. The importance of considering independence was demonstrated in section 2.5 using an artificial bivariate example.

[37] The performance of ICA and PCA as tools for the synthetic generation of rainfall and/or streamflow time series is now tested using a trivariate streamflow time series obtained from the Colombia data set described above. The trivariate case was selected to balance the need to induce a realistic level of complexity, while simultaneously acknowledging that higher-dimensional data sets would start to become troublesome given that the length of the streamflow time series is only 42 data points. To demonstrate the benefits of considering higher-order statistics, we adopt the approach illustrated in Figure 5, with each step described in more detail below:

[38] Step 1: We begin by decomposing the multivariate data set into univariate representations, using both PCA and ICA. As discussed above, the PCA solution involves diagonalizing the covariance matrix, while the ICA solution provides an additional rotation to the diagonalized data set so that an estimate of the mutual information is minimized.

[39] Step 2: We then apply a bootstrap with replacement to the individual PCs and ICs, thereby treating the components as univariate time series. Each bootstrapped sample is of length l, which is the length of the original component (in this case, 42 data points), and this was repeated to obtain p samples for each component, where p is set to 100 for the remainder of this analysis. The objective of the bootstrap is to maintain the marginal distribution of the components, while at the same time eliminating any joint dependence between the components which may be present.

[40] Step 3: The bootstrapped components are then rotated back to the original data space to obtain the synthetically generated PC and IC solutions, using the inverse of the rotation matrix obtained by PCA and ICA, respectively. Thus we have p synthetically generated, multivariate time series of length l from PCA and ICA. As highlighted by the example in section 2.5, the lack of independence can result in distortions to both the marginal and joint density attributes.

[41] Step 4: To support the visual assessment of performance described above, we use a kernel density estimate of both the joint density and the marginal densities to compare the performance of PCA and ICA. This is achieved by constructing a kernel density estimate using (1) the original multivariate data, (2) each of the p PCA-generated synthetic data sets, and (3) each of the p ICA-generated synthetic data sets. To ensure consistency between the density estimates, a grid was developed of size 50^d for the original multivariate data, where for this study we set $d = 3$ for the trivariate joint density, and $d = 1$ when considering the marginal densities, and this grid was used as the basis for all the density estimates. The bandwidth was estimated using the Gaussian reference bandwidth (for additional details, refer to *Sharma* [2000], and *Sharma et al.* [1998]).

[42] Step 5: To estimate the bias of the joint and marginal density estimates obtained from the PCA- and ICA-generated synthetic data sets compared to the joint and marginal density estimates of the original data, a kernel density surface was constructed as the unweighted mean of each of the p density estimates evaluated at each of the 50^d grid points.

Table 1. List of Three Streamflow Stations Used in the Analysis[a]

Station	Record	Statistic	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Mean
Alto Anchicaya	1962–2004	Mean	41.6	37.2	35.8	49.7	54.2	44.3	31.3	29.0	38.3	58.8	60.8	51.4	44.4
		Standard deviation	13.2	14.6	10.6	15.3	11.5	10.2	10.2	12.1	12.1	13.3	10.9	12.0	6.7
Rio Grande	1942–2004	Mean	23.3	22.0	23.6	32.4	40.0	36.6	32.6	32.8	36.5	44.5	43.6	32.6	33.5
		Standard deviation	7.5	8.6	8.9	10.1	11.2	10.4	10.5	10.4	11.1	11.1	9.4	8.5	6.6
Salvajina	1947–2004	Mean	165.2	143.9	136.5	149.8	151.1	127.0	103.8	74.8	63.2	109.5	195.7	212.4	136.8
		Standard deviation	67.2	69.7	64.7	51.9	47.3	34.6	26.7	17.7	21.4	43.0	66.1	86.7	31.5

[a]All units are m^3/s.

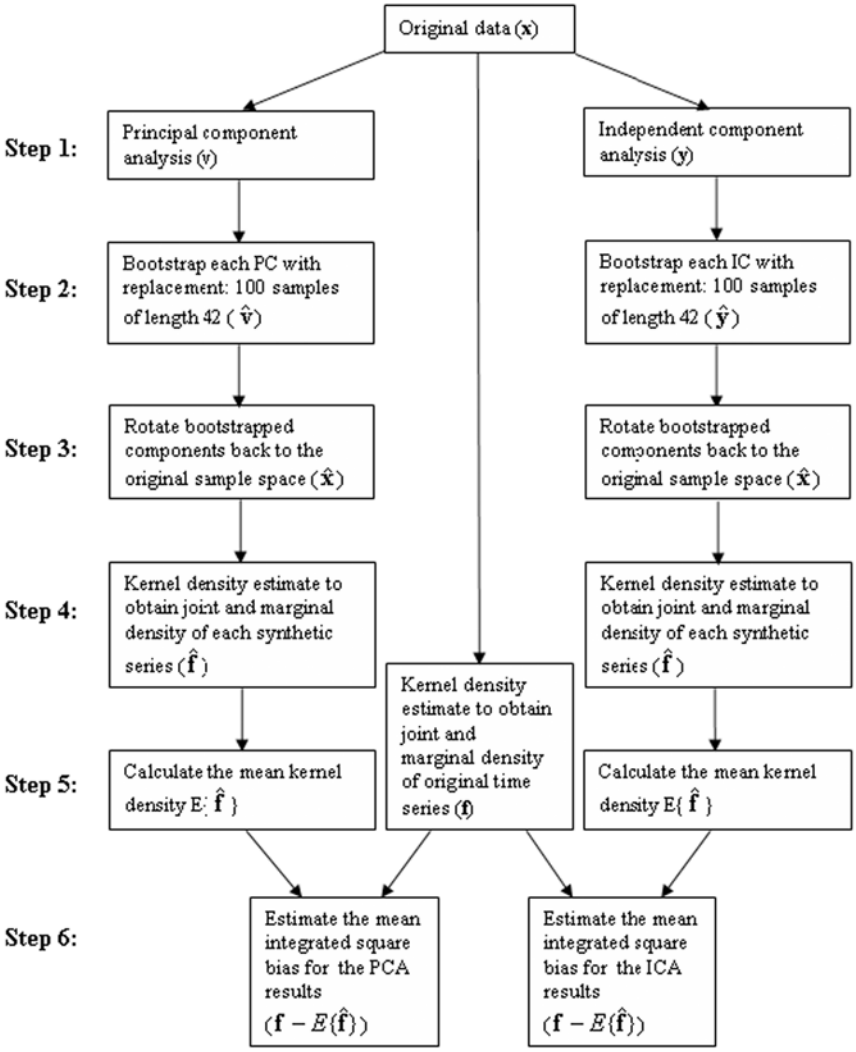

Figure 5. Stepwise approach to generating synthetic data using PCA and ICA.

[43] Step 6: The mean integrated squared bias (MISB) was then evaluated as follows [*Scott*, 1992]:

$$E\left\{\int (\mathbf{f} - E\{\hat{\mathbf{f}}\})d\mathbf{x}\right\} \tag{12}$$

where $E\{\}$ denotes the expectation operator over each of the bootstrap samples, $E\{\hat{\mathbf{f}}\}$ is the mean kernel density estimate from either the PCA or the ICA generated data sets, and f is the kernel density estimate of the original multivariate data, which is taken to be the true density. The MISB is calculated using both trivariate kernel density estimates to evaluate the joint dependence structure, and using univariate kernel density estimates to evaluate whether the original marginal distributions are maintained in the synthetic data.

[44] The above analysis allows for the comparison of PCA and ICA as a basis for generating synthetic data. This method is now tested on the multivariate streamflow data set described in the previous section.

4.1. Application to Three Streamflow Stations in Columbia

[45] To assist in visualizing the approach to streamflow generation, we demonstrate the technique using three streamflow time series from the Colombia data set, located at stations Rio Grande, Salvajina and Alto Anchicaya. This data set is presented in Figure 1, and can be considered to be a typical data set for most multivariate problems. Note that this data have been normalized, and therefore is centered around zero with a standard deviation of one.

[46] The results are shown in Figure 6. The data set is trivariate, although for clarity we present only the bivariate plots. The contours are therefore a trivariate density function integrated over the third (hidden) dimension, to form the bivariate contours shown.

[47] The left plots represent the original data set, with the kernel density estimate of this data superimposed. In the middle plots, the results from the ICA-derived data are shown, with the contours representing the mean kernel density estimate surface of the 100 bootstrapped samples. The same is repeated in the right plots, using the PCA results. In both the ICA and PCA plots, the original data points are superimposed on the contours.

[48] As can be seen from Figure 6, the ICA results show a much closer agreement to the original data compared with the PCA solution. In the top plots, where the stations Rio Grande and Salvajina are plotted against each other, it is apparent that one value is separated from the majority of

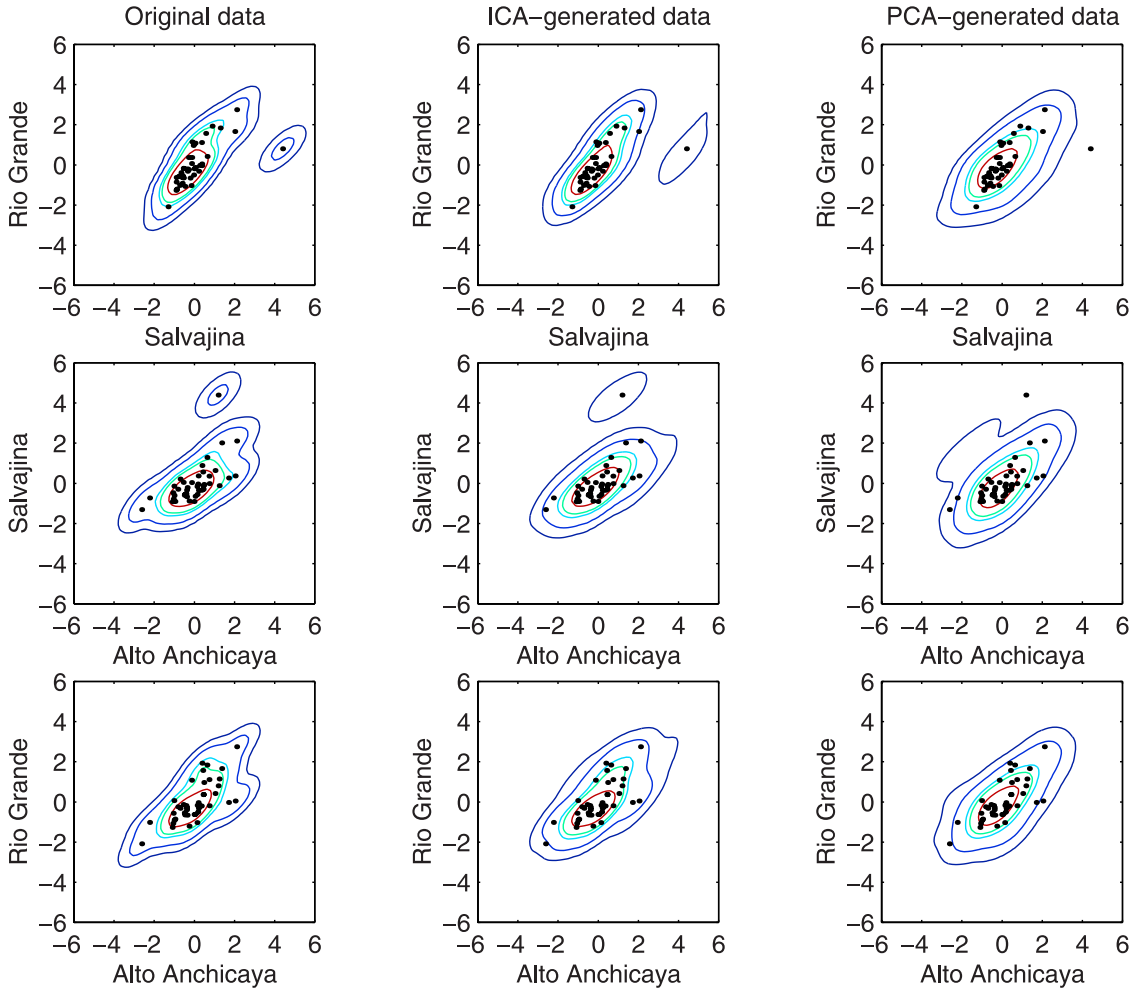

Figure 6. Results of the trivariate example first shown in Figure 1. (left) Original data with the trivariate kernel density estimate generated from the original data shown as contours. (middle) Original data but with the contours representing the kernel density estimate from the ICA-generated synthetic data. (right) Original data points with contours representing the kernel density estimate from the PCA-generated synthetic data.

other values. We assume this value has been properly recorded, and therefore does not constitute an outlier. As such, any stochastic generation technique should maintain the joint dependence implied by such a value. It is apparent that the ICA-derived data succeed in this attempt, while the PCA generated data are unable to regenerate this point. Furthermore, the contours of the PCA-derived data appear to be much further apart compared to both the ICA-derived data and the original data, indicating a redistribution of the probability across a broader region than was sampled in the historical record. A similar result is observed when examining the middle plots. Finally, in the bottom plots no such extreme values could be observed, although the dependence does not appear to be normally distributed. In particular, the variance of the conditional probability density of Rio Grande increase as Alto Anchicaya flows increase, which is better simulated by the ICA derived samples than for the PCA case. Readers should note that the joint probability plots for both ICA and PCA represent the average probability density across many realizations, and hence are smoother than similar plots across a single historical sample.

[49] Thus far, we have simply compared the visual performance of the time series. The MISB, described in section 2.5, was evaluated for this time series as well, and it was found to be 0.00228 for the ICA results, and 0.00652 for the PCA results. Thus the bias for the PCA results is approximately three times greater than for the ICA results. Improvements were also apparent in the estimation of the marginal distributions, with the marginal MISB for Rio Grande, Salvajina and Alto Anchicaya found to be 0.00019, 0.00033 and 0.00007 for the ICA analysis, and 0.00047, 0.00122 and 0.00011 for the PCA analysis, respectively.

4.2. General Application

[50] To test the consistency of the results, we considered the full range of trivariate data sets out of the 20 available streamflow records. In total, 1140 such combinations exist. One of the disadvantages to using ICA is that, unlike PCA, the solution requires an iterative approach to optimization, with associated problems of convergence and local optima. In the case of convergence, a total of 148 samples failed to converge, representing 13% of the total samples. We do not consider this to be a particular problem, however, because

Table 2. Comparison of the Mean Integrated Square Bias (MISB) When Using PCA and ICA, for the Univariate and Trivariate Cases

	MISB (PCA) Averaged Across All Sample Combinations	MISB (ICA) Averaged Across All Sample Combinations	Percentage Improvement, %
Marginal distribution	0.00061	0.00044	28
Trivariate distribution	0.00293	0.00219	25

nonconvergence is most likely to occur when the original data are close to a Gaussian distribution, since any orthogonal rotation of a Gaussian data set will remain Gaussian, the ICA solution would not be expected to improve significantly on the PCA solution [*Hyvarinen et al.*, 2001].

[51] The results therefore are based on the 992 trivariate combinations that converged, and are summarized in Table 2. The ICA MISB was between 27% and 176% of the PCA result, thereby suggesting a significant degree of variability between individual samples, with the ICA MISB being on average 25% lower than the PCA MISB when calculated over all 992 trivariate combinations. However, there were 170 cases (17% of the total sample) where the ICA MISB was higher than the PCA result. It is likely that the worsening performance for these stations was due to the difficulty in estimating statistical independence, particularly since the data set is only 42 data points long. A similar analysis was then conducted on the marginal distributions, and it was found that the MISB was on average about 28% lower for the ICA solution than for the PCA solution, although once again the results were highly variable.

[52] In contrast to the significant reduction in bias represented by the MISB, the sample variance did not change significantly between the ICA and PCA solutions, so that the mean integrated square error (MISE), a term which incorporates both the variance and bias components, was only 5% lower for ICA than for PCA. We hypothesize that the consistency in the variance results is due to the relationship between the variance component and the length of the bootstrap sample, which remains constant over the experiment.

5. Conclusions

[53] This paper presents a novel two-step approach for the synthetic generation of multivariate hydrological time series, which involves first decomposing the time series into univariate components, followed by synthetically generating additional data from the statistically independent, univariate time series, with the objective of maintaining the marginal probability density structure. The importance of maximizing independence was demonstrated through a comparison between PCA, a second-order method that diagonalizes the covariance matrix, and ICA which also considers higher-order statistics. As case studies, we considered first an artificial example based on a mixture of two independent, uniformly distributed samples, followed by real example which uses trivariate data drawn from a 20 dimensional data set of Colombian streamflow representing all major inflows to the hydroelectricity system (not presented).

[54] The results of this analysis of Colombian streamflow indicate that, on average, using ICA to decompose the multivariate data set results in an improvement in the manner in which joint dependence is represented, with the MISB of the joint being on average 25% lower than the PCA equivalent. Similarly, the MISB of the marginal distributions are on average 28% lower for the ICA solution when compared with the PCA solution. However, the ICA approach resulted in a higher MISB in a number of cases, which is mostly likely to be due to the difficulty in characterizing independence for sample lengths of 42 data points.

[55] This study therefore moves one step closer to developing a stochastic generation model that is able to simultaneously maintain spatial and temporal dependence. Areas for future research include applying a range of parametric and nonparametric univariate autoregressive models to the components to determine whether performance is maintained, and testing whether the method works at higher dimensions by incorporating a dimension reduction step in the analysis. A final area for future research involves comparing a range of alternative ICA algorithms such as algorithms that use direct estimates of mutual information [e.g., *Stogbauer et al.*, 2004], to determine the sensitivity of the results to the algorithm used. To the authors knowledge, this type of analysis has not been conducted for the relatively short data sets commonly used in hydrology.

Notation

d	dimension of the multivariate data set.
l	length of time series.
m	number of observations.
n	number of reduced dimensions.
p	number of bootstrap samples.
$\mathbf{x}$	an m-dimensional vector of observed data which has been centered.
$\mathbf{z}$	whitened data.
$\mathbf{s}$	an n-dimensional vector ofindependent components.
$\mathbf{y}$	estimator of $\mathbf{s}$.
$\mathbf{n}$	an m-dimensional vector of independent and identically distributed Gaussian noise.
$\mathbf{A}$	an $m \times n$ mixing matrix.
$\mathbf{w}_j$	weight vectors.
$\mathbf{W}$	weight matrix, $\mathbf{W} = (\mathbf{w_1}, \ldots, \mathbf{w}_n)^{\mathrm{T}}$.
$\mathbf{C_x}$	covariance matrix of $\mathbf{x}$, $\mathbf{C_x} = \boldsymbol{E}\{\mathbf{x}\mathbf{x}^{\mathrm{T}}\}$.
$\mathbf{e_i}$	eigenvectors of $\mathbf{C_x}$.
$\mathbf{E}$	matrix of eigenvectors.
$\mathbf{D}$	matrix of eigenvalues of $\mathbf{C_x}$, given as $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_n)$.
$\mathbf{PC}_i$	ith principal component of $\mathbf{x}$.
$\mathbf{V}$	whitening transform.
$\mathbf{I}$	identity matrix.
$E\{.\}$	mathematical expectation.
$f(.)$	probability density function.
$f_i(.)$	marginal probability density functions.
$G(.)$	a scalar nonlinear function.
$H(.)$	differential entropy.
$J(.)$	negentropy.

[56] **Acknowledgments.** The authors wish to thank Luis Fernando Puerta Correa from the Empresas Publicas de Medellin (EEPPM) for providing hydrologic data. Funding for this research came from the Australian Research Council and the Sydney Catchment Authority. Their support for this work is gratefully acknowledged.

References

Box, G. E. P., G. W. Jenkins, and G. C. Reinsel (1994), *Time Series Analysis—Forecasting and Control*, Prentice-Hall, Upper Saddle River, N. J.

Bras, R. L., and I. Rodrigues-Iturbe (1985), *Random Functions and Hydrology*, Addison-Wesley, Boston, Mass.

Comon, P. (1994), Independent component analysis: A new concept?, *Signal Process.*, *36*, 287–314.

Fraser, A. M., and H. L. Swinney (1986), Independent coordinates for strange attractors from mutual information, *Phys. Rev. A*, *33*, 1134–1140.

Golub, G. H., and C. F. Van Loan (1996), *Matrix Computations*, Johns Hopkins Univ. Press, Baltimore, Md.

Herault, J., and C. Jutten (1986), Space or time adaptive signal processing by neural network models, in *Neural Networks for Computing, AIP Conf. Proc.*, vol. 151, edited by J. S. Denker, pp. 206–211, Am. Inst. for Phys., New York.

Hyvarinen, A. (1997), One-unit contrast functions for independent component analysis: A statistical analysis, in *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, pp. 388–397, IEEE Press, Piscataway, N. J.

Hyvarinen, A. (1999), Survey on independent component analysis, *Neural Comput. Surv.*, *2*, 94–128.

Hyvarinen, A., and E. Oja (1997), A fast fixed-point algorithm for independent component analysis, *Neural Comput.*, *9*, 1483–1492.

Hyvarinen, A., and E. Oja (1998), Independent component analysis by general nonlinear Hebbian-like learning rules, *Signal Process.*, *64*, 301–313.

Hyvarinen, A., J. Karhunen, and E. Oja (2001), *Independent Component Analysis*, 481 pp., John Wiley, Hoboken, N. J.

Lall, U., and A. Sharma (1996), A nearest neighbour bootstrap for time series resampling, *Water Resour. Res.*, *32*, 679–693.

Lall, U., B. Rajagopalan, and D. G. Tarboton (1996), A nonparametric wet/dry spell model for resampling daily precipitation, *Water Resour. Res.*, *32*, 2803–2823.

Lee, T. W. (1998), *Independent Component Analysis—Theory and Applications*, Springer, New York.

Loucks, D. P., J. R. Stedinger, and D. A. Haith (1981), *Water Resource Systems Planning and Analysis*, Prentice-Hall, Upper Saddle River, N. J.

Oja, E. (2004), *Applications of Independent Component Analysis*, *Neural Inf. Process. Lecture Notes Comput. Sci.*, vol. 3316, Springer, Berlin.

Pegram, G. G. S., and W. James (1972), Multilag multivariate autoregressive model for the generation of operational hydrology, *Water Resour. Res.*, *8*, 1074–1076.

Salas, J. D. (1992), Analysis and modeling of hydrologic time series, in *Handbook of Hydrology*, edited by D. R. Maidment, pp. 19.1–19.72, McGraw-Hill, New York.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley, Hoboken, N. J.

Sharma, A. (2000), Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: part 3—A non-parametric probabilistic forecast model, *J. Hydrol.*, *239*, 249–258.

Sharma, A., and R. O'Neill (2002), A nonparametric approach for representing interannual dependence in monthly streamflow sequences, *Water Resour. Res.*, *38*(7), 1100, doi:10.1029/2001WR000953.

Sharma, A., D. G. Tarboton, and U. Lall (1997), Streamflow simulation: A nonparametric approach, *Water Resour. Res.*, *33*, 291–308.

Sharma, A., U. Lall, and D. G. Tarboton (1998), Kernel bandwidth selection for a first order nonparametric streamflow simulation model, *Stochastic Hydrol. Hydraul.*, *12*, 33–52.

Stogbauer, H., A. Kraskov, S. A. Astakhov, and P. Grassberger (2004), Least-dependent-component analysis based on mutual information, *Phys. Rev. E*, *70*, 066123.

Waymire, E., and V. K. Gupta (1981), The mathematical structure of rainfall representations: 1. A review of the stochastic rainfall models, *Water Resour. Res.*, *17*, 1261–1272.

Wilks, D. S. (1995), *Statistical Methods in the Atmospheric Sciences*, Elsevier, New York.

C. Brown and U. Lall, Department of Earth and Environmental Engineering, Columbia University, New York, NY 10027, USA. (caseyb@iri.columbia.edu; ula2@columbia.edu)

A. Sharma and S. Westra, School of Civil and Environmental Engineering, University of New South Wales, Sydney, NSW 2052, Australia. (a.sharma@unsw.edu.au; seth@civeng.unsw.edu.au)