Regression-Based Methods for Finding Coupled Patterns

MICHAEL K. TIPPETT

International Research Institute for Climate and Society, Palisades, New York

TIMOTHY DELSOLE

George Mason University, Fairfax, Virginia, and Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland

SIMON J. MASON AND ANTHONY G. BARNSTON

International Research Institute for Climate and Society, Palisades, New York

(Manuscript received 19 July 2007, in final form 29 January 2008)

ABSTRACT

There are a variety of multivariate statistical methods for analyzing the relations between two datasets. Two commonly used methods are canonical correlation analysis (CCA) and maximum covariance analysis (MCA), which find the projections of the data onto coupled patterns with maximum correlation and covariance, respectively. These projections are often used in linear prediction models. Redundancy analysis and principal predictor analysis construct projections that maximize the explained variance and the sum of squared correlations of regression models. This paper shows that the above pattern methods are equivalent to different diagonalizations of the regression between the two datasets. The different diagonalizations are computed using the singular value decomposition of the regression matrix developed using data that are suitably transformed for each method. This common framework for the pattern methods permits easy comparison of their properties. Principal component regression is shown to be a special case of CCA-based regression. A commonly used linear prediction model constructed from MCA patterns does not give a least squares estimate since correlations among MCA predictors are neglected. A variation, denoted least squares estimate (LSE)-MCA, is suggested that uses the same patterns but minimizes squared error. Since the different pattern methods correspond to diagonalizations of the same regression matrix, they all produce the same regression model when a complete set of patterns is used. Different prediction models are obtained when an incomplete set of patterns is used, with each method optimizing different properties of the regression. Some key points are illustrated in two idealized examples, and the methods are applied to statistical downscaling of rainfall over the northeast of Brazil.

1. Introduction

Multivariate statistical methods are used to analyze observational and model data, to make statistical forecasts, and to calibrate or correct dynamical forecasts. Some of the most commonly used methods include principal component analysis (PCA), maximum covariance analysis (MCA), and canonical correlation analysis (CCA; e.g., Bretherton et al. 1992). PCA is usually applied to a single dataset, finding the projections [empirical orthogonal functions (EOFs)] or components that explain the most variance. Methods such as CCA and MCA work with two datasets, finding projections that optimize some measure of linear association between the two datasets: CCA selects components of each dataset so as to maximize their correlation; MCA does likewise, except maximizing covariance. A common application of these methods is the construction of linear prediction models based on the identified, and often physically meaningful, coupled patterns.

Redundancy analysis (RDA) and principal predictor analysis (PPA) are pattern methods specifically tailored for use in linear regression models and, unlike CCA and MCA, are asymmetric in their treatment of the two

Corresponding author address: M. K. Tippett, International Research Institute for Climate and Society, The Earth Institute of Columbia University, Lamont Campus/61 Route 9W, Palisades, NY 10964.
E-mail: tippett@iri.columbia.edu

DOI: 10.1175/2008JCLI2150.1

datasets, identifying one dataset as the predictor and the other as the predictand. RDA selects predictor components that maximize explained variance (von Storch and Zwiers 1999; Wang and Zwiers 2001). PPA selects predictor components that maximize the sum of squared correlations (Thacker 1999). Another commonly used pattern regression method is principal component regression (PCR; e.g., Yu et al. 1997) in which PCA is applied to the predictor field and then a multiple linear regression is developed between the EOF coefficients or principal components (PCs) and each predictand element individually.

The purpose of this paper is to elucidate the connection between methods for finding coupled patterns and multivariate regression. A key element is the use of the singular value decomposition (SVD) to analyze the matrix of regression coefficients. The SVD reveals the structure of the regression by finding orthogonal transformations that diagonalize the regression. The singular values are the regression coefficients of the diagonalized regression. The regression is invariant with respect to linear transformations of the data (as long as the predictor transformation is invertible) in the sense that the regression matrix is transformed in the same way as the data. However, the SVD of the regression matrix is not invariant since, after a linear transformation of the data, the transformations that diagonalize the regression are generally no longer orthogonal. Therefore applying the SVD to regression matrices developed with different transformations of the data yields distinct diagonalizations of the regression. Furthermore, these distinct diagonalizations diagnose different properties of the regression as measured by the singular values. For instance, previous work has shown that when the data are expressed in the basis of its principal components, the regression matrix reduces to the cross-covariance matrix and its SVD corresponds to CCA with the singular values being the canonical correlations (Bretherton et al. 1992; DelSole and Chang 2003). Here we extend this idea and show that MCA, RDA, and PPA are equivalent to SVDs of the regression developed using data that are transformed in a distinct manner for each method. The connection between the pattern methods and multivariate regression provides a common framework that is useful for understanding and comparing the pattern methods, as well as for computation.

The paper is organized as follows: in section 2, we examine in a univariate regression how, with appropriate linear transformations of the data, the regression coefficient measures correlation, explained variance, or covariance. In section 3, we examine the behavior of multivariate regression when linear transformations are applied to the data. In section 4, we analyze the multivariate regression and obtain the coupled pattern methods as singular vectors of a transformed regression. We discuss reduced-rank regression in section 5. Some of the key issues are illustrated with idealized examples in section 6. The methods are compared in a statistical downscaling example in section 7. Section 8 gives a summary and conclusions.

2. Univariate linear regression

In the case of a single predictand and a single predictor, an estimate $\hat{y}$ of the predictand y based on the predictor x is given by the linear regression

$$\hat{y} = ax, \tag{1}$$

where the regression coefficient a is

$$a = \frac{\langle xy \rangle}{\langle x^2 \rangle}, \tag{2}$$

and $\langle\ \rangle$ denotes expectation; we take x and y to be anomaly values, that is, deviations from their respective means, and thus the regression equation contains no constant term. The regression coefficient can be manipulated to obtain quantities such as correlation, explained variance, and covariance, which measure aspects of the linear relation between predictor and predictand. Specifically,

$$\text{correlation} = \frac{\langle xy \rangle}{\sqrt{\langle x^2 \rangle \langle y^2 \rangle}} = a\sqrt{\frac{\langle x^2 \rangle}{\langle y^2 \rangle}},$$

$$\sqrt{\text{explained variance}} = \frac{|\langle xy \rangle|}{\sqrt{\langle x^2 \rangle}} = |a|\sqrt{\langle x^2 \rangle},$$

$$\text{covariance} = \langle yx \rangle = a\langle x^2 \rangle. \tag{3}$$

Here "explained variance" means the variance of y explained by the regression, *not* the fraction of variance, which is the square of the correlation. The difference of the variance of y and the explained variance is the error variance of the regression. Since the linear regression minimizes squared error, it maximizes explained variance.

The regression coefficient a changes in a simple way when a linear scaling is applied to the variables. Suppose that new variables are defined by $x' = lx$ and $y' = my$, where l and m are scalars and $l \neq 0$. The regression equation for the new variables is

$$\hat{y}' = a'x', \tag{4}$$

where the new regression coefficient a' is related to the original regression coefficient by

$$a' = \frac{\langle x'y' \rangle}{\langle x'^2 \rangle} = \frac{m}{l}\frac{\langle xy \rangle}{\langle x^2 \rangle} = \frac{m}{l}a. \tag{5}$$

Combining Eqs. (3) and (5) shows that particular choices of l and m lead to the transformed regression coefficient a' having the following interpretations:

- when both variables are normalized to have unit variance, $x' = x/\sqrt{\langle x^2 \rangle}$, $y' = y/\sqrt{\langle y^2 \rangle}$, and the regression coefficient a' is the correlation between x and y;
- when x alone is normalized to have unit variance, $x' = x/\sqrt{\langle x^2 \rangle}$, and the magnitude of the regression coefficient $|a'|$ is the square root of the variance explained by x; and
- when x is normalized by its variance, $x' = x/\langle x^2 \rangle$, and the regression coefficient a' is the covariance between x and y.

The connection between transformations of the data and the interpretation of the regression coefficient is simple but not particularly useful in the scalar case. The univariate regression does, however, indicate that rescaling of the data, while changing the value and interpretation of the regression coefficient, does not fundamentally change the regression; the rescalings of the data are simply applied to the least squares estimate (LSE) and the regression coefficient. This concept is generalized to the multivariate case in section 3, and in section 4 we present the appropriate multivariate generalizations of these data transformations that lead to regression coefficients that measure correlation, explained variance, or covariance—the same quantities that arise in methods for finding coupled patterns.

3. Multivariate linear regression

Suppose that the multivariate predictand y is linearly related to the multivariate predictor $\mathbf{x}$, where $\mathbf{x}$ and $\mathbf{y}$ are anomaly fields; we use the convention that $\mathbf{x}$ and $\mathbf{y}$ are column vectors. The least squares estimate $\hat{\mathbf{y}}$ of the predictand is given by linear regression as

$$\hat{\mathbf{y}} = \langle \mathbf{y}\mathbf{x}^{\mathrm{T}} \rangle \langle \mathbf{x}\mathbf{x}^{\mathrm{T}} \rangle^{-1}\mathbf{x}, \tag{6}$$

where T and -1 denote transpose and matrix inverse, respectively. Typically the expectations are computed from data using sample averages. The predictor data matrix $\mathbf{X}$ is the matrix whose ith column is the ith sample of the predictor $\mathbf{x}$; the number of rows of $\mathbf{X}$ is equal to the dimension of $\mathbf{x}$, and the number of columns of $\mathbf{X}$ is equal to the number of samples. Likewise the predictand data matrix $\mathbf{Y}$ is the matrix whose ith column is the ith sample of the predictand $\mathbf{y}$. Then

$$\hat{\mathbf{y}} = \mathbf{A}\mathbf{x}, \tag{7}$$

where the least squares regression coefficient matrix is defined as $\mathbf{A} \equiv (\mathbf{Y}\mathbf{X}^{\mathrm{T}})(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}$.[1]

As in the univariate case of Eq. (5), linear transformations of the data lead to transformation of the regression matrix. Suppose we introduce new variables $\mathbf{y}' = \mathbf{M}\mathbf{y}$ and $\mathbf{x}' = \mathbf{L}\mathbf{x}$, where $\mathbf{L}$ and $\mathbf{M}$ are matrices. The regression matrix $\mathbf{A}'$ relating the transformed variables is

$$\mathbf{A}' = \mathbf{Y}'\mathbf{X}'^{\mathrm{T}}(\mathbf{X}'\mathbf{X}'^{\mathrm{T}})^{-1} = (\mathbf{M}\mathbf{Y}\mathbf{X}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}})(\mathbf{L}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}})^{-1}. \tag{8}$$

If, additionally, $\mathbf{L}$ is invertible then the transformed regression matrix has the simple form

$$\mathbf{A}' = \mathbf{M}\mathbf{A}\mathbf{L}^{-1}, \tag{9}$$

analogous to the univariate case in Eq. (5). This relation provides several pieces of useful information. First, when the transformation $\mathbf{L}$ of the predictor is invertible, the least squares estimate $\hat{\mathbf{y}}'$ of $\mathbf{y}'$ is

$$\hat{\mathbf{y}}' = \mathbf{A}'\mathbf{x}' = \mathbf{M}\mathbf{A}\mathbf{L}^{-1}\mathbf{L}\mathbf{x} = \mathbf{M}\hat{\mathbf{y}}, \tag{10}$$

which means that the least squares estimate using the transformed data is just the transformation of the original least squares estimate. Rescaling the data or expressing it in another basis does not fundamentally change the regression so long as the transformation $\mathbf{L}$ of predictor data is invertible.

The transformation $\mathbf{L}$ of the predictor data is not invertible when $\mathbf{L}\mathbf{x} = 0$ for some $\mathbf{x} \neq 0$, which means that the transform $\mathbf{L}$ has the effect of reducing the number of predictors. Reducing the number of predictors is often desirable when the dimension of the predictor is large compared to the number of available samples. When the number of predictors is large compared to the number of samples, the sample covariance matrix $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ is ill conditioned or even singular, and reducing the number of predictors regularizes the regression problem by making it have a unique solution that is not overly sensitive to the data. The number of predictors is often reduced using PCA, which finds the components of the data that explain the most variance, although other projections may be used as well (DelSole and Shukla 2006). Reducing the set of predictors to some smaller number of PCs is called *prefiltering* in the con-

[1] We use the convention that the data in $\mathbf{X}$ and $\mathbf{Y}$ are normalized by $\sqrt{n-1}$, where n is the number of samples. This convention simplifies the notation by making $\langle \mathbf{x}\mathbf{x}^{\mathrm{T}} \rangle = \mathbf{X}\mathbf{X}^{\mathrm{T}}$ and $\langle \mathbf{y}\mathbf{x}^{\mathrm{T}} \rangle = \mathbf{Y}\mathbf{X}^{\mathrm{T}}$. The matrix of regression coefficients is independent of n.

text of CCA (Bretherton et al. 1992). In contrast to the general case of singular transformations of the predictor data, when $\mathbf{L}$ is the prefiltering transformation that maps the data onto a subset of its PCs, the regression developed with the prefiltered data is the same as the original regression applied to the prefiltered data (see the appendix).

The goal when selecting the number of predictors is a skillful model. However, the data used to estimate the regression coefficients are not directly useful for determining the skill of the regression.[2] For instance, if the dimension of the predictor $\mathbf{x}$ exceeds the number of samples and prefiltering is done using the maximum number of PCs, the in-sample error is 0 since $\mathbf{Y} = \mathbf{AX}$. However, since such a regression completely fits the data, including its random components, we expect it to suffer from *overfitting* and have poor skill on independent data. Regression models with fewer predictors are more likely to represent the actual relationships, avoid overfitting, and better predict out-of-sample data. To choose the number of predictors that optimizes the out-of-sample skill of the regression, the data can be split into two segments with the regression coefficients estimated using one segment and the number of predictors chosen to optimize the skill in the independent segment. This procedure does, however, give an overly optimistic estimate of skill due to selection bias (Zucchini 2000), and the skill of the selected model should ideally be estimated on a third independent set of data. In what follows, we assume that the number of predictors has been reduced so that the number of predictors is less than the number of samples, and the predictor covariance matrix is invertible.

Another important consequence of the relation in Eq. (8) follows from noting that the error variance $\|\mathbf{y}' - \hat{\mathbf{y}}\|^2 = (\mathbf{y}' - \hat{\mathbf{y}}')^{\mathrm{T}}(\mathbf{y} - \hat{\mathbf{y}}')$ of the transformed variable is minimized, and that $(\mathbf{y}' - \hat{\mathbf{y}}')^{\mathrm{T}}(\mathbf{y}' - \hat{\mathbf{y}}') = (\mathbf{y} - \hat{\mathbf{y}})^{\mathrm{T}}(\mathbf{M}^{\mathrm{T}}\mathbf{M})\,(\mathbf{y} - \hat{\mathbf{y}})$. Therefore, not only is the sum of squared error $(\mathbf{y} - \hat{\mathbf{y}})^{\mathrm{T}}(\mathbf{y} - \hat{\mathbf{y}})$ minimized, but so is the positive semidefinite quadratic function of the error $(\mathbf{y} - \hat{\mathbf{y}})^{\mathrm{T}}(\mathbf{M}^{\mathrm{T}}\mathbf{M})(\mathbf{y} - \hat{\mathbf{y}})$. Changing the weighting of the error elements does not result in an estimate that is different from the least squares estimate, and in fact, some of the weights can be set to 0 since $\mathbf{M}$ is not required to be invertible. For instance, choosing $\mathbf{M} = \mathbf{e}_i^{\mathrm{T}}$, where $\mathbf{e}_i$ is the ith column of the identity matrix means that the least squares estimate minimizes $\|\mathbf{e}_i^{\mathrm{T}}(\mathbf{y} - \hat{\mathbf{y}})\|^2 = (y_i - \hat{y}_i)^2$, which is the error of the ith element of the predictand. Therefore regression minimizes not only the total error variance but the error variance of each element separately. Consequently, the regression estimate developed using all the elements of the predictand simultaneously is the same as the one developed with individual elements of the predictand separately. However, for questions of inference, such as testing hypotheses about the regression coefficients, the multivariate character of the problem cannot be neglected, and correlations between parameters must be considered.

This last property of regression aids the interpretation of PCR. In PCR, regressions are developed between predictor PCs and each of the predictands individually. The above conclusion means that PCR is the same as developing the regression between all of the predictands and the PCs simultaneously. This shows a connection with CCA since a CCA-based regression model with EOF prefiltering of the predictor (and no other truncation) is the same as multiple linear regression between the predictor PCs and the predictand (Glahn 1968; DelSole and Chang 2003). Therefore PCR is the same as a CCA-based regression model with EOF prefiltering of the predictor and no other truncation such as prefiltering of the predictand.

4. Analysis of the regression matrix

We now show that transforming the multivariate data in ways suggested by the univariate case allows us to interpret the regression coefficients as correlation, variance explained, standardized explained variance, or covariance of the original data. The SVD of the transformed regression matrix diagonalizes the regression and identifies projections of the data that maximize these measures. These projections are the same that are used in methods for finding coupled patterns.

a. Correlation

In the univariate case, normalizing the predictor and predictand by their standard deviation makes the regression coefficient equal to the correlation between predictor and predictand. The appropriate multivariate generalization is to multiply the variables by the inverse of the matrix square root[3] of their covariance:

$$\mathbf{x}' = (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1/2}\mathbf{x}$$

$$\mathbf{y}' = (\mathbf{Y}\mathbf{Y}^{\mathrm{T}})^{-1/2}\mathbf{y}. \tag{11}$$

[2] There are in-sample estimates of the out-of-sample error such as Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) that take into account the number of predictors (Akaike 1973; Schwarz 1978).

[3] A matrix square root of the positive definite matrix $\mathbf{P}$ is $\mathbf{Z}$ if $\mathbf{Z}\mathbf{Z}^{\mathrm{T}} = \mathbf{P}$.

The appearance of the inverse of the predictand covariance indicates that it may be necessary to prefilter the predictand as well as the predictor. The matrix square root is not uniquely defined; postmultiplication of a matrix square root by an orthogonal matrix gives another matrix square root. A convenient choice for the matrix square root of the inverse covariance is the transformation that replaces the data with its PC time series (normalized to have unit variance); that is, $\mathbf{x}'$ and $\mathbf{y}'$ are the normalized principal component scores. Such a transformation is sometimes called a *whitening* transformation (DelSole and Tippett 2007) since the transformed data are uncorrelated and have unit variance

$$\mathbf{X}'\mathbf{X}'^{\mathrm{T}} = \mathbf{I} \quad \text{and} \quad \mathbf{Y}'\mathbf{Y}'^{\mathrm{T}} = \mathbf{I}, \tag{12}$$

where $\mathbf{I}$ is the identity matrix. The regression matrix for predicting $\mathbf{y}'$ from $\mathbf{x}'$ is

$$\mathbf{A}' = \mathbf{Y}'\mathbf{X}'^{\mathrm{T}}(\mathbf{X}'\mathbf{X}'^{\mathrm{T}})^{-1} = \mathbf{Y}'\mathbf{X}'^{\mathrm{T}}, \tag{13}$$

since $\mathbf{X}'\mathbf{X}'^{\mathrm{T}} = \mathbf{I}$. The (i, j)th element of $\mathbf{A}'$ is the correlation between the ith element of $\mathbf{y}'$ and the jth element of $\mathbf{x}'$, denoted y_i' and x_j', respectively, since

$$\mathbf{A}'_{ij} = \mathbf{e}_i^{\mathrm{T}}\mathbf{A}'\mathbf{e}_j = \mathbf{e}_i^{\mathrm{T}}\mathbf{Y}'\mathbf{X}'^{\mathrm{T}}\mathbf{e}_j = \mathbf{e}_i^{\mathrm{T}}\mathbf{Y}'(\mathbf{e}_j^{\mathrm{T}}\mathbf{X}')^{\mathrm{T}} = \langle y_i' x_j' \rangle, \tag{14}$$

and the elements of $\mathbf{x}'$ and $\mathbf{y}'$ have unit variance.

Instead of looking at the correlations between individual elements of $\mathbf{x}'$ and $\mathbf{y}'$, we can examine the correlation of one-dimensional projections of the data. Projecting the transformed predictand and predictor data onto the *weight vectors* $\mathbf{u}$ and $\mathbf{v}$, respectively, gives the time series

$$\frac{\mathbf{u}^{\mathrm{T}}\mathbf{Y}'}{\sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{u}}} \quad \text{and} \quad \frac{\mathbf{v}^{\mathrm{T}}\mathbf{X}'}{\sqrt{\mathbf{v}^{\mathrm{T}}\mathbf{v}'}}, \tag{15}$$

which from Eq. (12) have unit variance. The correlation between the time series of the projections is

$$\frac{\mathbf{u}^{\mathrm{T}}\mathbf{Y}'(\mathbf{v}^{\mathrm{T}}\mathbf{X}')^{\mathrm{T}}}{\sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{u}\mathbf{v}^{\mathrm{T}}\mathbf{v}}} = \frac{\mathbf{u}^{\mathrm{T}}\mathbf{Y}'\mathbf{X}'^{\mathrm{T}}\mathbf{v}}{\sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{u}\mathbf{v}^{\mathrm{T}}\mathbf{v}}} = \frac{\mathbf{u}^{\mathrm{T}}\mathbf{A}'\mathbf{v}}{\sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{u}\mathbf{v}^{\mathrm{T}}\mathbf{v}}}, \tag{16}$$

where we use the definition of $\mathbf{A}'$ from Eq. (13). This ratio is maximized when $\mathbf{u}$ and $\mathbf{v}$ are, respectively, the left and right leading singular vectors of $\mathbf{A}'$ (Golub and Van Loan 1996). The SVD of $\mathbf{A}'$ is defined to be

$$\mathbf{A}' = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}, \tag{17}$$

where $\mathbf{U}$ and $\mathbf{V}$ are square orthogonal matrices and $\mathbf{S}$ is a matrix with nonnegative diagonal entries s_i ordered from largest to smallest; the columns of $\mathbf{U}$ and $\mathbf{V}$ form complete, orthogonal bases for the predictand and predictor, respectively. The singular vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ are the ith columns of $\mathbf{U}$ and $\mathbf{V}$ and satisfy

$$\mathbf{u}_i^{\mathrm{T}}\mathbf{A}'\mathbf{v}_i = s_i, \quad i = 1, \ldots, k, \tag{18}$$

where k is the smaller of row and column dimensions of $\mathbf{A}'$. Therefore $s_1 = \mathbf{u}_1^{\mathrm{T}}\mathbf{A}'\mathbf{v}_1$ is the largest possible correlation between projections of the data. The next largest singular value $s_2 = \mathbf{u}_2^{\mathrm{T}}\mathbf{A}'\mathbf{v}_2$ is the largest possible correlation between projections of the data subject to the constraint that the projections be orthogonal to the first ones, that is, the constraint that $\mathbf{u}_2^{\mathrm{T}}\mathbf{u}_1 = \mathbf{v}_2^{\mathrm{T}}\mathbf{v}_1 = 0$. This orthogonality constraint has the consequence that the associated time series are uncorrelated because

$$(\mathbf{u}_1^{\mathrm{T}}\mathbf{Y}')(\mathbf{u}_2^{\mathrm{T}}\mathbf{Y}')^{\mathrm{T}} = \mathbf{u}_1^{\mathrm{T}}\mathbf{u}_2 = 0, \quad \text{and}$$
$$(\mathbf{v}_1^{\mathrm{T}}\mathbf{X}')(\mathbf{v}_2^{\mathrm{T}}\mathbf{X}')^{\mathrm{T}} = \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_2 = 0. \tag{19}$$

Likewise, subsequent singular values are the maximum correlation subject to the constraint that the projections are orthogonal (time series are uncorrelated) to previous ones.

The weight vectors for the untransformed variables are the columns of the matrices $\mathbf{Q}_x$ and $\mathbf{Q}_y$ defined so that the projection of the untransformed variables is equal to the projection of the transformed variables

$$\mathbf{Q}_x^{\mathrm{T}}\mathbf{X} = \mathbf{V}^{\mathrm{T}}\mathbf{X}' \quad \text{and} \quad \mathbf{Q}_y^{\mathrm{T}}\mathbf{Y} = \mathbf{U}^{\mathrm{T}}\mathbf{Y}'. \tag{20}$$

Using Eq. (11) gives $\mathbf{Q}_y = (\mathbf{Y}\mathbf{Y}^{\mathrm{T}})^{-1/2}\mathbf{U}$ and $\mathbf{Q}_x = (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1/2}\mathbf{V}$. Although the weight vectors for the transformed variables are orthogonal, the weight vectors for the untransformed variables are not, or more precisely, they are orthogonal with respect to a different norm since $\mathbf{Q}_y^{\mathrm{T}}(\mathbf{Y}\mathbf{Y}^{\mathrm{T}})^{-1}\mathbf{Q}_y = \mathbf{I}$ and $\mathbf{Q}_x^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{Q}_x = \mathbf{I}$. The data can be expressed as *patterns* that multiply the time series. The pattern vectors differ from the weight vectors since the weight vectors are not orthogonal. The matrices $\mathbf{P}_x$ and $\mathbf{P}_y$ of pattern vectors are found by solving

$$\mathbf{X} = \mathbf{P}_x\mathbf{Q}_x^{\mathrm{T}}\mathbf{X} \quad \text{and} \quad \mathbf{Y} = \mathbf{P}_y\mathbf{Q}_y^{\mathrm{T}}\mathbf{Y}, \tag{21}$$

which gives $\mathbf{P}_x = (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{1/2}\mathbf{V}$ and $\mathbf{P}_y = (\mathbf{Y}\mathbf{Y}^{\mathrm{T}})^{1/2}\mathbf{U}$; these patterns solve Eq. (21) in a least squares sense when an incomplete set of projections is used. The pattern and weight vectors are orthogonal to each other since $\mathbf{P}_x^{\mathrm{T}}\mathbf{Q}_x = \mathbf{I}$ and $\mathbf{P}_y^{\mathrm{T}}\mathbf{Q}_y = \mathbf{I}$.

The above analysis defines the decomposition of the data into patterns whose times series have the maximum correlation subject to the constraint that subsequent predictor and predictand time series be uncorrelated. This decomposition is CCA with the columns of $\mathbf{Q}_x(\mathbf{Q}_y)$ being the predictor (predictand) weight vectors, the columns of $\mathbf{P}_x(\mathbf{P}_y)$ the predictor (predictand) patterns, and the diagonal elements of $\mathbf{S}$ the canonical correlations [DelSole and Chang (2003); see the appen-

Table 1. The optimized quantity, the variable transformations, the weights, and the patterns for CCA, RDA, PPA, and MCA. In all cases, $\mathbf{USV}^{\mathrm{T}}$ is the SVD of the transformed regression $\mathbf{A}' = \mathbf{Y}'\mathbf{X}'^{\mathrm{T}}$ and the decomposition of the original regression is $\mathbf{A} = \mathbf{P}_y\mathbf{S}\mathbf{Q}_x^{\mathrm{T}}$.

	CCA	RDA	PPA	MCA
Optimizes	Correlation	Explained variance	Sum of squared correlations	Covariance
$\mathbf{x}'$	$\mathbf{x}' = (\mathbf{XX}^{\mathrm{T}})^{-1/2}\mathbf{x}$	$\mathbf{x}' = (\mathbf{XX}^{\mathrm{T}})^{-1/2}\mathbf{x}$	$\mathbf{x}' = (\mathbf{XX}^{\mathrm{T}})^{-1/2}\mathbf{x}$	$\mathbf{x}' = (\mathbf{XX}^{\mathrm{T}})^{-1}\mathbf{x}$
$\mathbf{y}'$	$\mathbf{y}' = (\mathbf{YY}^{\mathrm{T}})^{-1/2}\mathbf{y}$	$\mathbf{y}' = \mathbf{y}$	$\mathbf{y}' = (\mathrm{Diag}\ \mathbf{YY}^{\mathrm{T}})^{-1/2}\mathbf{y}$	$\mathbf{y} = \mathbf{y}'$
Weights	$\mathbf{Q}_x = (\mathbf{XX}^{\mathrm{T}})^{-1/2}\mathbf{V}$	$\mathbf{Q}_x = (\mathbf{XX}^{\mathrm{T}})^{-1/2}\mathbf{V}$	$\mathbf{Q}_x = (\mathbf{XX}^{\mathrm{T}})^{-1/2}\mathbf{V}$	$\mathbf{Q}_x = (\mathbf{XX}^{\mathrm{T}})^{-1/2}\mathbf{V}$
	$\mathbf{Q}_y = (\mathbf{YY}^{\mathrm{T}})^{-1/2}\mathbf{U}$	$\mathbf{Q}_y = \mathbf{U}$	$\mathbf{Q}_y = (\mathrm{Diag}\ \mathbf{YY}^{\mathrm{T}})^{-1/2}\mathbf{U}$	$\mathbf{Q}_y = \mathbf{U}$
Patterns	$\mathbf{P}_x = (\mathbf{XX}^{\mathrm{T}})^{1/2}\mathbf{V}$	$\mathbf{P}_x = (\mathbf{XX}^{\mathrm{T}})^{1/2}\mathbf{V}$	$\mathbf{P}_x = (\mathbf{XX}^{\mathrm{T}})^{1/2}\mathbf{V}$	$\mathbf{P}_x = (\mathbf{XX}^{\mathrm{T}})^{1/2}\mathbf{V}$
	$\mathbf{P}_y = (\mathbf{YY}^{\mathrm{T}})^{1/2}\mathbf{U}$	$\mathbf{P}_y = \mathbf{U}$	$\mathbf{P}_y = (\mathrm{Diag}\ \mathbf{YY}^{\mathrm{T}})^{1/2}\mathbf{U}$	$\mathbf{P}_y = \mathbf{U}$

dix of this paper for a derivation of the usual CCA equations]. Using the relation between $\mathbf{A}$ and $\mathbf{A}'$ in Eq. (9), the regression matrix can be simply written using the weight vectors and patterns as

$$\mathbf{A} = (\mathbf{YY}^{\mathrm{T}})^{1/2}\mathbf{A}'(\mathbf{XX}^{\mathrm{T}})^{-1/2} = (\mathbf{YY}^{\mathrm{T}})^{1/2}\mathbf{USV}^{\mathrm{T}}(\mathbf{XX}^{\mathrm{T}})^{-1/2} = \mathbf{P}_y\mathbf{SQ}_x^{\mathrm{T}}. \quad (22)$$

The above relation shows that CCA diagonalizes the regression. Since $\mathbf{Q}_x^{\mathrm{T}}\mathbf{P}_x = \mathbf{I}$, $\mathbf{AP}_x = \mathbf{P}_y\mathbf{S}$, and predictor patterns are mapped to predictand patterns scaled by their correlation. The decomposition of $\mathbf{A}$ in Eq. (22) is not the usual SVD of $\mathbf{A}$ since $\mathbf{P}_y$ and $\mathbf{Q}_x$ are not orthogonal matrices, but can be interpreted as a SVD of $\mathbf{A}$ with the usual vector norms replaced by the norms implied by the whitening transformations.[4]

b. Variance explained

In the univariate case, normalizing the predictor by its standard deviation and leaving the predictand unchanged makes the regression coefficient equal the square root of the explained variance. The appropriate generalization to the multivariate problem is to apply the whitening transformation to the predictor as in Eq. (11),

$$\mathbf{x}' = (\mathbf{XX}^{\mathrm{T}})^{-1/2}\mathbf{x}, \quad (23)$$

and to leave the predictand unchanged. The regression matrix relating $\mathbf{x}'$ and $\mathbf{y}$ is

$$\mathbf{A}' = \mathbf{YX}'^{\mathrm{T}}, \quad (24)$$

since $\mathbf{X}'\mathbf{X}'^{\mathrm{T}} = \mathbf{I}$. Proceeding as in (14) of the previous section shows that the absolute value of the (i, j) entry of the transformed regression matrix $\mathbf{A}'$ is the square root of variance explained by the regression between y_i and x'_j. The square root of the variance explained by a regression between projections using the weight vectors $\mathbf{u}$ and $\mathbf{v}$ of the predictand and predictor data is

$$\frac{\mathbf{u}^{\mathrm{T}}\mathbf{Y}(\mathbf{v}^{\mathrm{T}}\mathbf{X}')^{\mathrm{T}}}{\sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{u}\mathbf{v}^{\mathrm{T}}\mathbf{v}}} = \frac{\mathbf{u}^{\mathrm{T}}\mathbf{A}'\mathbf{v}}{\sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{u}\mathbf{v}^{\mathrm{T}}\mathbf{v}}}. \quad (25)$$

This ratio is maximized when $\mathbf{u}$ and $\mathbf{v}$ are, respectively, the leading left and right singular vectors of $\mathbf{A}'$. Therefore $s_1^2 = (\mathbf{u}_1^{\mathrm{T}}\mathbf{A}'\mathbf{v}_1)^2$ is the maximum variance explained by a single predictor. Conversely, $\langle \mathbf{y}^{\mathrm{T}}\mathbf{y} \rangle - s_1^2$ is the minimum error variance of a regression that uses a single predictor. The variance explained using the first two pairs of singular vectors is $s_1^2 + s_2^2$, and the minimum error variance when two predictors are used is $\langle \mathbf{y}^{\mathrm{T}}\mathbf{y} \rangle - s_1^2 - s_2^2$. The variances add since the predictor projections are uncorrelated, a consequence of the fact that $\mathbf{v}_1^{\mathrm{T}}\mathbf{X}'\mathbf{X}'^{\mathrm{T}}\mathbf{v}_2 = \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_2 = 0$. The predictand projection time series are correlated but the predictand weight (and pattern) vectors are orthogonal since $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_2 = 0$. This decomposition of the data is called RDA (von Storch and Zwiers 1999; Wang and Zwiers 2001). Additional details of the weight and pattern vectors are given in Table 1. The RDA patterns diagonalize the regression with the diagonal elements measuring the square root of the variance explained by each predictor pattern. A related method is Empirical Orthogonal Teleconnection 2 (EOT2), which finds the predictor element, rather than the linear combination of predictor elements, that explains the maximum predictand variance (Van den Dool 2006). Subsequent uncorrelated EOT2 components are computed iteratively by finding the predictor element at each step that explains the most residual variance.

c. Explained standardized variance

If the variances of the predictands are highly disparate, *standardization,* that is, normalizing each predictand by its standard deviation, may be appropriate. Applying RDA to standardized predictands finds the projections that maximize the explained standardized variance. Explicitly, we use the transformations

[4] The dependence of the SVD on choice of norm is well known in ensemble forecasting where the SVD is sometimes used to generate initial perturbations (Ehrendorfer and Tribbia 1997).

$$\mathbf{x}' = (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1/2}\mathbf{x}$$

$$\mathbf{y}' = (\mathrm{Diag}\ \mathbf{Y}\mathbf{Y}^{\mathrm{T}})^{-1/2}\mathbf{y}, \tag{26}$$

where the notation Diag $\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$ means the diagonal matrix whose diagonal elements are the same as those of $\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$; the elements of the diagonal matrix Diag $\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$ are the predictor variances and $\mathbf{y}'$ is $\mathbf{y}$ with each element divided by its standard deviation. This transformation of the predictand normalizes each predictand to have unit variance as in CCA, but unlike CCA, the transformed predictands remain correlated. The transformed regression matrix is

$$\mathbf{A}' = \mathbf{Y}'\mathbf{X}'^{\mathrm{T}}. \tag{27}$$

The (i, j)th element of $\mathbf{A}'$ is the correlation between y_i and x_j' since

$$\mathbf{A}'_{ij} = \mathbf{e}_i^{\mathrm{T}}\mathbf{A}'\mathbf{e}_j = \mathbf{e}_i^{\mathrm{T}}\mathbf{Y}'\mathbf{X}'^{\mathrm{T}}\mathbf{e}_j = \frac{1}{\sqrt{\mathbf{e}_i^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{e}_i}}\,\mathbf{e}_i^{\mathrm{T}}\mathbf{Y}(\mathbf{e}_j^{\mathrm{T}}\mathbf{X}')^{\mathrm{T}}. \tag{28}$$

The absolute value of this quantity is also the square root of the fraction of the variance of y_i explained by x_j', that is, the square root of the explained standardized variance. Paralleling the interpretation of CCA and RDA, we project the transformed data onto the vectors $\mathbf{u}$ and $\mathbf{v}$. The square root of the standardized explained variance of the regression between the projections is

$$\frac{\mathbf{u}^{\mathrm{T}}\mathbf{Y}'(\mathbf{v}^{\mathrm{T}}\mathbf{X}')^{\mathrm{T}}}{\sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{u}\mathbf{v}^{\mathrm{T}}\mathbf{v}}} = \frac{\mathbf{u}^{\mathrm{T}}\mathbf{A}'\mathbf{v}}{\sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{u}\mathbf{v}^{\mathrm{T}}\mathbf{v}}}. \tag{29}$$

The $\mathbf{u}$ and $\mathbf{v}$ that maximize this ratio are the leading singular vectors of $\mathbf{A}'$.

The explained standardized variance is the sum of the explained fraction of variance for each predictand. On the other hand, the explained fraction of variance for each predictand is the square of the correlation between the prediction and the predictand. Therefore, maximizing the explained standardized variance is the same as maximizing the sum of squared correlations between predictand and prediction.

We call this decomposition of data PPA after Thacker (1999) who focused on the predictor patterns, which he called principal predictors and characterized as maximizing the sum of squared correlation between the predictor patterns and the predictand data. Like CCA and RDA, the PPA predictor projections are uncorrelated because of the use of the whitening transformation. However, the predictand projections are neither uncorrelated nor orthogonal. Additional details of the weight and pattern vectors are give in Table 1. PPA provides a diagonalization of the regression with the diagonal elements measuring the square root of the explained standardized variance for each pattern pair.

d. Covariance

In the univariate problem, normalizing the predictor by its variance makes the regression coefficient equal to covariance. To generalize to the multivariate problem we multiply the predictor by the inverse of its covariance,

$$\mathbf{x}' = (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{x}, \tag{30}$$

and do not transform $\mathbf{y}$. The regression matrix for predicting $\mathbf{y}$ from $\mathbf{x}'$ is

$$\mathbf{A}' = \mathbf{Y}\mathbf{X}'^{\mathrm{T}}(\mathbf{X}'\mathbf{X}'^{\mathrm{T}})^{-1} = \mathbf{Y}\mathbf{X}^{\mathrm{T}}. \tag{31}$$

The (i, j)th element of $\mathbf{A}'$ is the covariance between y_i and x_j. The covariance between projections of the predictand and predictor data in the directions $\mathbf{u}$ and $\mathbf{v}$, respectively, is

$$\frac{\mathbf{u}^{\mathrm{T}}\mathbf{Y}(\mathbf{v}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}}{\sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{u}\mathbf{v}^{\mathrm{T}}\mathbf{v}}} = \frac{\mathbf{u}^{\mathrm{T}}\mathbf{A}'\mathbf{v}}{\sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{u}\mathbf{v}^{\mathrm{T}}\mathbf{v}}}. \tag{32}$$

This ratio is maximized when $\mathbf{u}$ and $\mathbf{v}$ are the left and right leading singular vectors of $\mathbf{A}'$.[5] This decomposition of the data is maximum covariance analysis (MCA), sometimes referred to as SVD; we use the name MCA to distinguish between the coupled pattern method based on the SVD of the cross covariance and the general SVD matrix procedure (von Storch and Zwiers 1999).

Writing the regression matrix $\mathbf{A}$ using the MCA projections gives

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}, \tag{33}$$

where $\mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$ is the SVD of $\mathbf{Y}\mathbf{X}^{\mathrm{T}}$.[6] The ith MCA predictand projection is uncorrelated with the jth MCA predictor projection for $i \neq j$ since the matrix $\mathbf{U}^{\mathrm{T}}\mathbf{Y}\mathbf{X}^{\mathrm{T}}\mathbf{V}$ only has nonzero elements on its diagonal. However, this does not mean that a regression for the ith MCA predictand projection should only include the ith MCA predictor projection. To show this, we express the pre-

[5] The SVD of the cross-covariance matrix also arises in the solution of the orthogonal Procrustes problem (Gower and Dijksterhuis 2004).

[6] Forming $\mathbf{Y}\mathbf{X}^{\mathrm{T}}$ is impractical and unnecessary when the predictor and predictand dimensions are large compared to the number of samples. Instead, MCA can be applied to the covariance of the unnormalized predictor and predictand PCs since the SVD is invariant under orthogonal transformations. The dimensions in the SVD calculation are thus determined by the number of PCs rather than the predictor and predictand dimensions.

dictor data as $\mathbf{X} = \mathbf{VB}$, where the rows of $\mathbf{B}$ contain the time series of the projection of the predictors onto $\mathbf{V}$, and $\mathbf{B}$ is given by $\mathbf{B} = \mathbf{V}^{\mathrm{T}}\mathbf{X}$. Substituting this representation of the predictor data into Eq. (33) gives

$$\mathbf{A} = (\mathbf{USV}^{\mathrm{T}})(\mathbf{VBB}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}})^{-1} = \mathbf{US}(\mathbf{BB}^{\mathrm{T}})^{-1}\mathbf{V}^{\mathrm{T}}. \quad (34)$$

This form is similar to usual MCA-based linear models. However, usually $\mathbf{BB}^{\mathrm{T}}$ is replaced by the diagonal matrix whose first n_y diagonal entries are the same as those of $\mathbf{BB}^{\mathrm{T}}$, the variance of the MCA time series and whose remaining diagonal entries are zero (e.g., Widmann et al. 2003); n_y is the dimension of the predictand. This approximation means that correlations between MCA modes are neglected, and the resulting estimate is not generally an LSE. Therefore, we call the method using the regression matrix in Eq. (34) LSE-MCA since it uses the projections that maximize covariance like MCA but *is* a least squares estimate. Like MCA, LSE-MCA requires no EOF prefiltering. Widmann (2005) also noted that the usual MCA-based linear models do not agree with CCA-based regression and multiple linear regression, even when the predictand is a scalar. When the predictand is a scalar, the usual MCA-based linear model uses a single SVD mode as predictor, truncating the predictor data. LSE-MCA, on the other hand, does not truncate the data and reproduces the least squares estimate. The usual MCA-based linear model is the same as LSE-MCA when the predictor and predictand dimensions are the same and $\mathbf{BB}^{\mathrm{T}}$ is indeed diagonal. The matrix $\mathbf{BB}^{\mathrm{T}}$ is diagonal when the MCA modes $\mathbf{V}$ also happen to be EOFs of the predictor or when the predictors are uncorrelated with equal variance and the covariance matrix $\mathbf{XX}^{\mathrm{T}}$ is proportional to the identity matrix; the latter condition is true when, for instance, the predictors are whitened variables. Feddersen et al. (1999) used MCA projections in a least squares estimate but with an implementation that additionally required the solution be found by numerical optimization. MCA is similar to *partial least squares* (PLS) regression (Wold et al. 1984; Boulesteix and Strimmer 2007) in that the components maximize covariance, and the first PLS component is the same as the first MCA component. However, subsequent components differ because PLS components are uncorrelated.

5. Reduced-rank regressions

We have shown that the regression matrix can be decomposed into patterns that optimize selected quantities including correlation, explained variance, explained standardized variance, and covariance. These decompositions help diagnose properties of the regression by expressing the data in bases so that the regression matrix is diagonal. As shown in section 3, the use of different bases does not fundamentally change the regression as long as the bases are complete and there is no truncation of the data. Therefore, all the methods give the same prediction model as multiple regression when a complete set of patterns are used. However, the regression is changed when a partial set of patterns is used, effectively truncating the data used to develop the regression. Such a simplification of the regression may be desirable since it reduces the number of predictors, and hence the number of parameters that must be estimated from the data. We expect that regressions that use too many predictors will have poor skill on independent data due to overfitting and sampling error.

Reducing the number of patterns used in the regression is somewhat different from prefiltering, which reduces the number of predictors or predictands without necessarily considering joint relations between predictor and predictand. Decomposition of the regression into pairs of patterns produces measures of the strength of the relation between the patterns; for instance, CCA gives the correlation between the time series of the patterns. Therefore, it is reasonable to retain those pairs of patterns that represent the strongest relations and discard the rest. Since overfitting may exaggerate the in-sample relationship, validation of the relation on independent data is useful for deciding which pairs of patterns to retain. Often cross-validated skill is the basis for selecting the patterns to keep in the regression. However, as mentioned earlier in the context of EOF prefiltering, the cross-validated skill used to select the model will give an overly optimistic estimate of performance on independent data due to selection bias.

Since the pattern pairs are found by computing the SVD of the transformed regression matrix $\mathbf{A}'$, restricting the patterns used in the regression is the same as replacing $\mathbf{A}'$ by a truncated SVD; that is, the regression matrix $\mathbf{A}' = \mathbf{USV}^{\mathrm{T}}$ is replaced with $\dot{\mathbf{A}}' = \mathbf{U}\dot{\mathbf{S}}\mathbf{V}^{\mathrm{T}}$, where the first r diagonal elements of $\dot{\mathbf{S}}$ are the same as those of $\mathbf{S}$ and the rest are 0; the patterns and weights do not change. The resulting truncated regression matrix $\dot{\mathbf{A}} = \mathbf{P}_y\dot{\mathbf{S}}\mathbf{Q}_x^{\mathrm{T}}$ retains r pairs of patterns and has the property that it is the rank-r regression, which optimizes the condition that the SVD measures. In particular, depending on method, the rank-r regression may optimize mutual information (CCA),[7] explained variance (RDA), the

[7] This is a consequence of the facts that (i) mutual information of normally distributed variables is an increasing function of correlation alone and (ii) the mutual information of a sum of independent variables is the sum of their mutual information.

sum of squared correlations (PPA), or the sum of covariance (LSE-MCA).

The patterns obtained in each method are generally different, and for a given value of r, the rank-r regression will be different for each method. Therefore, the different pattern methods produce different regressions when the regression is truncated. The motives of the user or the nature of the problem may indicate that one pattern method is preferable over another. For instance, CCA can select patterns with large correlation but small explained variance. In this case, RDA might be preferable as it maximizes explained variance. Similarly, LSE-MCA, by maximizing covariance, may select patterns with large variance but not necessarily high correlation. In this case, CCA might be preferable. The optimization of mutual information makes CCA attractive from the viewpoint of predictability since mutual information is a predictability measure with many attractive properties (DelSole and Tippett 2007).

A final point regarding these truncated regressions is that the truncated regression is indeed the same as the regression developed using the data projected onto the retained patterns since essentially a diagonal regression matrix is being truncated.

6. Two idealized examples

a. MCA and LSE-MCA

We now consider a simple example that illustrates the difference between the commonly used MCA linear model and LSE-MCA. We take $\mathbf{x}$ and $\mathbf{y}$ each to have two elements. Suppose that $\mathbf{YX}^{\mathrm{T}} = \mathbf{I}$ and the MCA modes are the columns of the identity matrix. Then from Eq. (33) the regression matrix is simply

$$\mathbf{A} = (\mathbf{XX}^{\mathrm{T}})^{-1}. \tag{35}$$

The commonly used approximation of the regression matrix is the diagonal matrix

$$\mathbf{A}_{\mathrm{MCA}} = \mathrm{Diag}(\mathbf{XX}^{\mathrm{T}})^{-1}. \tag{36}$$

We take the predictor covariance to have the form

$$\mathbf{XX}^{\mathrm{T}} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \sigma^2 \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}^{\mathrm{T}}, \tag{37}$$

where θ is the angle between the MCA modes and predictor EOFs, and the predictor EOFs have variance of 1 and σ^2. The angle θ is important because MCA and LSE-MCA are the same when the MCA modes are predictor EOFs, that is, when $\theta = 0$. Additionally, suppose the predictand covariance has the similar structure

$$\mathbf{YY}^{\mathrm{T}} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} 1/(0.8)^2 & 0 \\ 0 & 1/(0.5\sigma)^2 \end{bmatrix} \times \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}^{\mathrm{T}}. \tag{38}$$

These choices for the covariances imply that the CCA weight vectors are the same as the EOFs and that the canonical correlations are 0.8 and 0.5 [see Eq. (A5) of the appendix].

The error variance of the regression is

$$\begin{aligned} \langle \|\mathbf{Ax} - \mathbf{y}\|^2 \rangle &= \mathrm{tr}(\mathbf{A}\langle \mathbf{xx}^{\mathrm{T}} \rangle \mathbf{A}^{\mathrm{T}} + \langle \mathbf{yy}^{\mathrm{T}} \rangle - \mathbf{A}\langle \mathbf{xy}^{\mathrm{T}} \rangle - \langle \mathbf{yx}^{\mathrm{T}} \rangle \mathbf{A}^{\mathrm{T}}) \\ &= \mathrm{tr}(\mathbf{AXX}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} + \mathbf{YY}^{\mathrm{T}} - \mathbf{A} - \mathbf{A}^{\mathrm{T}}) \\ &= \mathrm{tr}[\mathbf{YY}^{\mathrm{T}} - (\mathbf{XX}^{\mathrm{T}})^{-1}] \\ &= \frac{1}{0.8^2} + \frac{1}{(0.5\sigma)^2} - 1 - \frac{1}{\sigma^2}, \end{aligned} \tag{39}$$

where we use the facts that $\mathbf{YX}^{\mathrm{T}} = \mathbf{I}$ and $\mathbf{A} = (\mathbf{XX}^{\mathrm{T}})^{-1}$. The error variance of the MCA model is

$$\begin{aligned} \langle \|\mathbf{A}_{\mathrm{MCA}}\mathbf{x} - \mathbf{y}\|^2 \rangle &= \mathrm{tr}(\mathbf{A}_{\mathrm{MCA}}\mathbf{XX}^{\mathrm{T}}\mathbf{A}_{\mathrm{MCA}}^{\mathrm{T}} + \mathbf{YY}^{\mathrm{T}} - \mathbf{A}_{\mathrm{MCA}} - \mathbf{A}_{\mathrm{MCA}}^{\mathrm{T}}) \\ &= \mathrm{tr}(\mathbf{YY}^{\mathrm{T}} - (\mathbf{XX}^{\mathrm{T}})^{-1}) + \mathrm{tr}(\mathbf{A}_{\mathrm{MCA}}\mathbf{XX}^{\mathrm{T}}\mathbf{A}_{\mathrm{MCA}}^{\mathrm{T}} - (\mathbf{XX}^{\mathrm{T}})^{-1}) \\ &= \langle \|\mathbf{Ax} - \mathbf{y}\|^2 \rangle + \mathrm{tr}(\mathbf{A}_{\mathrm{MCA}}\mathbf{XX}^{\mathrm{T}}\mathbf{A}_{\mathrm{MCA}}^{\mathrm{T}} - (\mathbf{XX}^{\mathrm{T}})^{-1}). \end{aligned} \tag{40}$$

Therefore the error variance of the MCA linear model relative to that of the LSE-MCA regression is

$$\frac{\langle \|\mathbf{A}_{\mathrm{MCA}}\mathbf{x} - \mathbf{y}\|^2 \rangle}{\langle \|\mathbf{Ax} - \mathbf{y}\|^2 \rangle} = 1 + \frac{\mathrm{tr}(\mathbf{A}_{\mathrm{MCA}}\mathbf{XX}^{\mathrm{T}}\mathbf{A}_{\mathrm{MCA}}^{\mathrm{T}} - (\mathbf{XX}^{\mathrm{T}})^{-1})}{\langle \|\mathbf{Ax} - \mathbf{y}\|^2 \rangle}. \tag{41}$$

This error variance is governed by θ and σ. Figure 1 shows the error of the MCA linear model relative to that of the LSE-MCA regression as a function of θ and σ. When $\theta = 0$, the MCA modes are also EOFs of the predictors, and there is no difference between the methods. Increasing θ increases the error of the MCA linear model. When $\sigma = 1$, the methods are the same since $\mathbf{XX}^{\mathrm{T}} = \mathbf{I}$ and again the MCA modes are the same as the predictor EOFs. As σ decreases, the relative error of the MCA linear model increases.

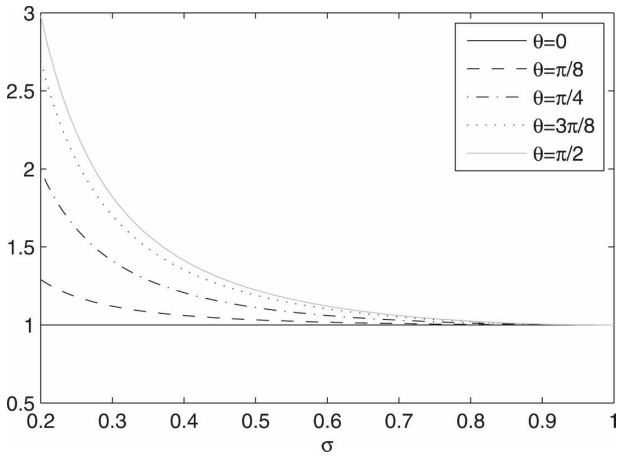

FIG. 1. Ratio of the MCA linear model error to that of the LSE-MCA regression as a function of σ for different values of the angle θ between predictor EOFs and MCA modes (see text).

b. CCA, LSE-MCA, and RDA

We now present a simple example to illustrate some issues regarding the truncation of the regression as discussed in section 5. We construct a two-dimensional, diagonal example where the correlations of the two elements are specified and examine the error of rank-1 regressions as the variance of one of the elements is varied. In particular, suppose that

$$\mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{Y}\mathbf{Y}^{\mathrm{T}} = \begin{bmatrix} 1 & 0 \\ 0 & \sigma^2 \end{bmatrix}. \tag{42}$$

The first and second elements are uncorrelated and have variance 1 and σ^2, respectively, for both $\mathbf{x}$ and $\mathbf{y}$. Note that σ^2 may or may not exceed 1. Suppose that $\mathbf{Y}\mathbf{X}^{\mathrm{T}}$ is diagonal and given by

$$\mathbf{Y}\mathbf{X}^{\mathrm{T}} = \begin{bmatrix} c_1 & 0 \\ 0 & c_2\sigma^2 \end{bmatrix}, \tag{43}$$

so that c_1 and c_2 are the canonical correlations; $c_1 \geq c_2$. The regression matrix is

$$\mathbf{A} = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix}, \tag{44}$$

and the regression error variance is

$$\langle \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 \rangle = (1 - c_1^2) + (1 - c_2^2)\sigma^2. \tag{45}$$

The rank-1 CCA regression selects the part of the system with highest correlation, which is the first element, regardless of σ. We now show that when the first element has little variance, the regression based on the leading CCA pattern does not minimize squared error. The rank-1 CCA regression matrix is

$$\mathbf{A}_1^{\mathrm{CCA}} = \begin{bmatrix} c_1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{46}$$

The error variance of the rank-1 CCA regression relative to the full regression is

$$\frac{1 - c_1^2 + \sigma^2}{(1 - c_1^2) + (1 - c_2^2)\sigma^2}. \tag{47}$$

On the other hand, LSE-MCA selects the part of the system with highest covariance. Examination of Eq. (43) shows that for $\sigma < \sqrt{c_1/c_2}$, the first element has the highest covariance and the rank-1 LSE-MCA regression matrix is the same as the rank-1 CCA regression matrix. For $\sigma > \sqrt{c_1/c_2}$, the second element has highest covariance and the rank-1 LSE-MCA regression matrix is

$$\mathbf{A}_1^{\text{LSE-MCA}} = \begin{bmatrix} 0 & 0 \\ 0 & c_2 \end{bmatrix}, \tag{48}$$

and the error variance relative to the full regression is

$$\frac{1 + (1 - c_2^2)\sigma^2}{(1 - c_1^2) + (1 - c_2^2)\sigma^2}. \tag{49}$$

Comparing Eq. (47) with Eq. (49) shows that the error variance of the rank-1 LSE-MCA regression is larger than that of the rank-1 CCA regression when $\sqrt{c_1/c_2} \leq \sigma \leq c_1/c_2$ and smaller when $\sigma \geq c_1/c_2$. This result agrees with the intuition that if σ is small, we expect the squared error to be minimized by the rank-1 regression matrix accounting for the first element, which has highest correlation. On the other hand, if σ is sufficiently large, then the rank-1 regression should be based on the second element.

Figure 2 shows the squared error of the rank-1 regressions as a function of σ for $c_1 = 0.8$ and $c_2 = 0.5$. There are three regimes. For $\sigma \leq \sqrt{c_1/c_2} \approx 1.26$, the LSE-MCA and CCA rank-1 regressions are the same. For $\sqrt{c_1/c_2} \leq \sigma \leq c_1/c_2$, the error of the LSE-MCA rank-1 regression is greater than that of the rank-1 CCA regression because the LSE-MCA is selecting the second element since it explains more covariance. However, the second element has lower correlation, and the resulting regression has higher rms error. For $\sigma \geq c_1/c_2 = 1.6$, the error of the rank-1 CCA regression is larger than that of the rank-1 LSE-MCA regression because the large value of σ dominates the rms error. In this simple two-dimensional example, the RDA rank-1 regression coincides with either the CCA or LSE-MCA rank-1 regressions, depending on which one has smaller

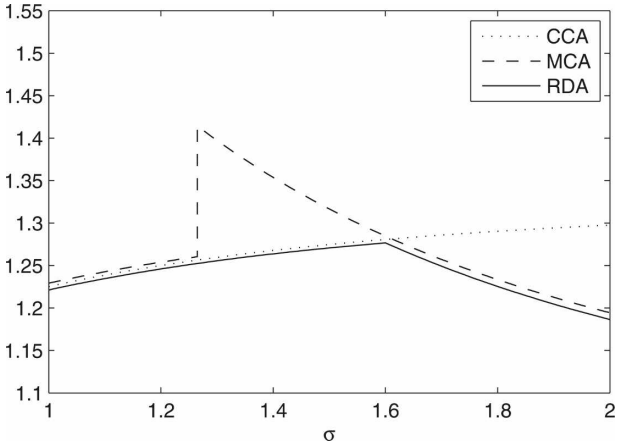

FIG. 2. Error of the rank-1 regression relative to that of the full regression as a function of σ for $c_1 = 0.8$ and $c_2 = 0.5$. Curves are offset for legibility.

rms error. In general, RDA-based regression is distinct from and has smaller rms error than either CCA or MCA-based regressions of the same rank.

7. Example: Statistical downscaling

General circulation models (GCMs) often have relatively coarse horizontal spatial resolution. Information about smaller scales can sometimes be extracted from the coarse-scale GCM output by forming a regression between GCM output and observations (Widmann et al. 2003). Such a regression can also be used to remove systematic model errors (Feddersen et al. 1999). We apply this procedure to the ensemble mean of a 24-member set of ECHAM4.5 (Roeckner et al. 1996) T42 ensemble simulations of March–May (1950–2000) precipitation over the northeast of Brazil (13°S–1°N, 55°–35°W). With T42 model resolution corresponding to a 2.8° × 2.8° spatial grid, the model domain contains 63 grid points. During this time of the year, precipitation over the northeast of Brazil is closely related to sea surface temperature (SST), and the GCM forced with observed SST skillfully reproduces some aspects of seasonal precipitation interannual variability. Observational data are taken from a gridded (0.5° × 0.5°) rainfall observation dataset (New et al. 2000). Leave-one-out cross validation is used to select the level of EOF prefiltering[8] as well as the number of patterns retained in the regression; the truncations for each method are chosen to maximize the sum over grid points of those cross-validated correlations greater than 0.3. Results using rms error as a truncation metric are similar, though rms error tends to select lower-dimensional models, as has been noted generally (Browne 2000).

The correlation map for a cross-validated univariate per gridpoint regression between the gridded observations and the GCM output interpolated to the observation grid is shown in Fig. 3a. Although there is a large region with correlations greater than 0.5, the gridpoint regression is limited by not using spatial correlation information. Figures 3b–g show the correlation maps of regressions based on PCR, CCA, RDA, MCA, LSE-MCA, and PPA patterns, respectively. All of the spatial pattern regression methods show overall improvement compared to the gridpoint regression and are fairly similar to each other. Their similarity may reflect that there seems to be only one or two meaningful modes in the regression that are captured by all the methods. The CCA regression uses five predictor EOFs and two predictand EOFs to form a rank-2 regression; RDA and PPA use rank-1 regressions based on five and three predictor EOFs, respectively.

The best overall results for correlation skill are obtained with CCA. Although CCA is expected to perform better than PCR since PCR is the special case of CCA with an untruncated predictand, there is no particular reason to expect CCA to outperform LSE-MCA or RDA, in general. The differences in skill are mostly insignificant, in a statistical sense. Both MCA and LSE-MCA use the same four modes that maximize covariance. Although we expect LSE-MCA to perform better than MCA since MCA neglects correlations between predictors, the impact of sampling error on the performance of the methods is unknown. One could imagine poor estimation of the correlations among the predictors outweighing neglecting interpredictor correlations. In any case, in this example, LSE-MCA does outperform MCA, which has the worst performance of the pattern regression methods. The regression with the smallest cross-validated rms error is RDA. The rank-1 CCA regression (not shown) has slightly lower overall correlation than the rank-2 CCA regression, but has lower cross-validated rms error, in fact, lower than that of the RDA regression.

8. Summary and conclusions

Two commonly used linear methods for finding coupled patterns in two datasets are canonical correlation analysis (CCA) and maximum covariance analysis (MCA), which find projections of the data having maximum correlation and covariance, respectively. Such methods are useful for diagnosing relations between

[8] Predictor and predictand are prefiltered in CCA. Only the predictor is prefiltered in RDA and PPA. No prefiltering is used with MCA or LSE-MCA.

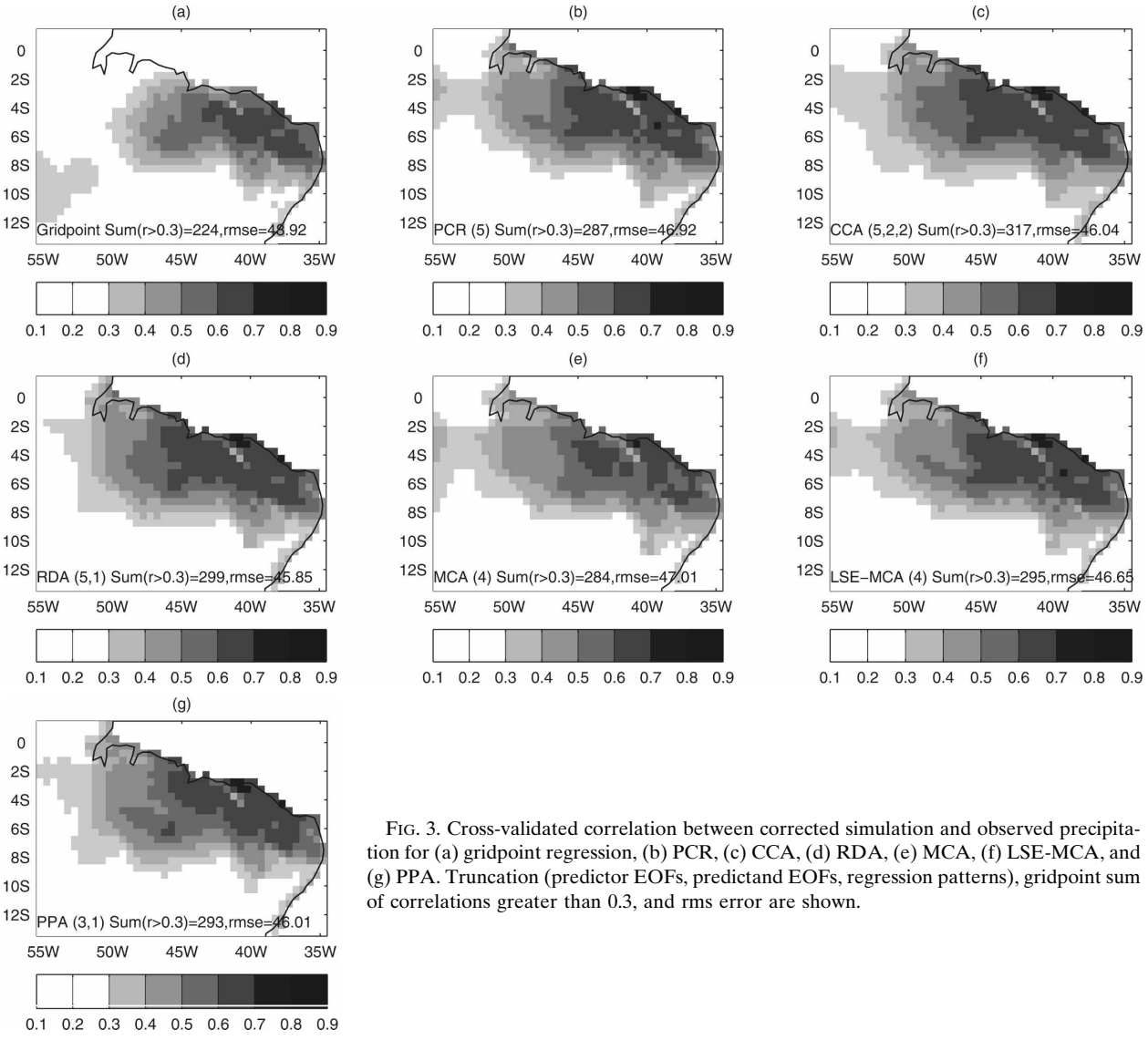

FIG. 3. Cross-validated correlation between corrected simulation and observed precipitation for (a) gridpoint regression, (b) PCR, (c) CCA, (d) RDA, (e) MCA, (f) LSE-MCA, and (g) PPA. Truncation (predictor EOFs, predictand EOFs, regression patterns), gridpoint sum of correlations greater than 0.3, and rms error are shown.

variables and constructing linear prediction models. Pattern methods like redundancy analysis (RDA) and principal predictor analysis (PPA) were developed specifically for use in linear prediction models and maximize explained variance and the sum of squared correlations, respectively. In this paper we show that these methods diagonalize the regression and are singular value decompositions (SVDs) of the matrix of regression coefficients for data transformed suitably for each respective method.

The essential character of the regression does not change when linear transformations are applied to data, as long as the transformation of the predictors is invertible. One consequence of the invariance of the regression is that regression-based prediction minimizes not only the sum of squared errors but any positive semidefinite quadratic function of the error. This fact implies that the regressions developed with each predictand individually will give the same predictions as the regression developed with all the predictands simultaneously. Consequently, principal component regression (PCR) in which regressions are developed between predictor PCs and individual predictands gives the same prediction model as does the regression developed between the set of predictands and the predictor PCs simultaneously, which in turn is the same as CCA with EOF prefiltering of the predictor and no other truncations.

Although the regression is invariant under linear transformations of the data, the meaning of the regression coefficients changes depending on the transformation of the data. This connection between the interpre-

tation of the regression coefficients and transformation of the data is readily apparent in the univariate case where differing normalizations of the data determine whether the regression coefficient measures correlation, explained variance, or covariance. Analogous transformations in the multivariate case lead to the regression matrix having coefficients that measure the same quantities. The whitening transformation in which the data are replaced by its normalized PCs is the multivariate generalization of normalizing a variable by its standard deviation.

The structure of the regression matrix is revealed by the SVD, which finds orthogonal bases so that the regression matrix is diagonal. Depending on the transformation applied to the data, the singular values measure correlation, explained variance, explained standardized variance, or covariance. The singular vectors identify the projections of the data that optimize these quantities and correspond to the methods CCA, RDA, PPA, and MCA, respectively. The SVD of a transformed regression can also be interpreted as the SVD of the untransformed regression with particular choices of norm for the predictor and predictand (Ehrendorfer and Tribbia 1997).

A common method for constructing a linear prediction model from MCA patterns does not produce a least squares estimate since correlations between MCA predictors are neglected. A variation, LSE-MCA, uses the same MCA patterns which maximize covariance but minimizes squared error. There are some special cases when MCA and LSE-MCA are the same, such as when the predictor and predictand dimensions are the same and MCA patterns are also EOFs of the predictor. In general, as illustrated in a two-dimensional example, the MCA linear model will have larger rms error than LSE-MCA. In practice, where sampling error plays a role, the MCA linear model may potentially gain some benefit by neglecting poorly estimated correlations among the predictors. However, in statistical downscaling GCM simulated rainfall over the northeast of Brazil, the MCA model had slightly worse performance compared to the other pattern methods.

Since the different coupled pattern methods correspond to decompositions of the same regression matrix, they all produce the same prediction model when a complete set of patterns is used. The choice of pattern method is important to the regression model when the SVD is truncated, that is, when an incomplete set of patterns is used. The regression model obtained by retaining only the first r pairs of patterns is the rank-r regression that maximizes mutual information, explained variance, explained standardized variance, and covariance for CCA, RDA, PPA, and LSE-MCA, respectively. We illustrate in a two-dimensional example that the RDA rank-1 regression is the rank-1 regression that minimizes rms error while the rank-1 regressions based on CCA or MCA patterns generally do not.

TABLE 2. CCA, RDA, and PPA expressed as MCA of transformed data.

$$\mathrm{CCA}(\mathbf{X}, \mathbf{Y}) = \mathrm{MCA}[(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1/2}\mathbf{X}, (\mathbf{Y}\mathbf{Y}^{\mathrm{T}})^{-1/2}\mathbf{Y}]$$
$$\mathrm{RDA}(\mathbf{X}, \mathbf{Y}) = \mathrm{MCA}[(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1/2}\mathbf{X}, \mathbf{Y}]$$
$$\mathrm{PPA}(\mathbf{X}, \mathbf{Y}) = \mathrm{MCA}[(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1/2}\mathbf{X}, (\mathrm{Diag}\ \mathbf{Y}\mathbf{Y}^{\mathrm{T}})^{-1/2}\mathbf{Y}]$$

The difference between reduced-rank regressions based on the different methods depends on the difference between the subspaces spanned by the retained patterns of each method, not differences between individual patterns. For instance, although the first r RDA patterns (assuming $r > 1$) may be different from the first r CCA patterns, if they collectively span the same subspace, regressions based on them will be identical. This fact may help in understanding why all the methods produce linear models with comparable skill in the statistical downscaling example.

The derivation of the pattern methods in the regression framework makes it easy to compare the methods and is useful for computation. A practical benefit of this approach is that an algorithm or computational method developed for one method is easily adapted for the other methods by transforming the data. For instance, Table 2 shows that all the methods can be expressed as MCA applied to transformed data.

An important issue that has not been examined closely here is the role of sampling error. The finite number of samples causes sampling error to affect all the methods, such that the underlying covariances are imperfectly known. EOF prefiltering is only one method for limiting the covariances to information that can be robustly estimated. Ridge methods are another approach to treat this problem (Vinod 1976; Hastie et al. 1995).

Acknowledgments. We thank two anonymous reviewers for their useful comments and suggestions. We thank Benno Blumenthal for the IRI Data Library. This paper was funded in part by a grant/cooperative agreement from the National Oceanic and Atmospheric Administration (NOAA), contract NA07GP0213 with the Trustees of Columbia University. The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its subagencies. The second author was supported by the National Science Foundation (ATM0332910), National Aeronautics and Space Administration (NNG04GG46G),

and the National Oceanographic and Atmospheric Administration (NA04OAR4310034).

APPENDIX

EOF Prefiltering and CCA Equations

a. EOF prefiltering

Let $\mathbf{L} = \mathbf{Z}^{\mathrm{T}}$, where $\mathbf{Z}$ is a matrix whose columns contain some but not all of the orthogonal eigenvectors of the predictor covariance matrix $\mathbf{X}\mathbf{X}^{\mathrm{T}}$. Then $\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{Z} = \mathbf{Z}\boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix containing the corresponding eigenvalues. The regression matrix relating $\mathbf{y}$ and $\mathbf{x}' = \mathbf{L}\mathbf{x}$ is

$$\begin{aligned}\mathbf{A}' &= (\mathbf{Y}\mathbf{X}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}})(\mathbf{L}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}})^{-1}\\ &= (\mathbf{Y}\mathbf{X}^{\mathrm{T}}\mathbf{Z})(\mathbf{Z}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{Z})^{-1}\\ &= (\mathbf{Y}\mathbf{X}^{\mathrm{T}}\mathbf{Z})(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\boldsymbol{\Lambda})^{-1}\\ &= \mathbf{Y}\mathbf{X}^{\mathrm{T}}\mathbf{Z}\boldsymbol{\Lambda}^{-1}. \qquad \text{(A1)}\end{aligned}$$

The projection $\mathbf{P}$ that projects the predictor data onto the space spanned by the columns of $\mathbf{Z}$ is $\mathbf{P} = \mathbf{Z}\mathbf{Z}^{\mathrm{T}}$. Applying the original regression to the projected data is the same as the regression with the transformed data because

$$\begin{aligned}\mathbf{A}\mathbf{P}\mathbf{x} &= \mathbf{A}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{x}\\ &= (\mathbf{Y}\mathbf{X}^{\mathrm{T}})(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{x}\\ &= (\mathbf{Y}\mathbf{X}^{\mathrm{T}})\mathbf{Z}\boldsymbol{\Lambda}^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{x}\\ &= \mathbf{A}'\mathbf{L}\mathbf{x}\\ &= \mathbf{A}'\mathbf{x}'. \qquad \text{(A2)}\end{aligned}$$

b. Alternative form for CCA

The usual CCA equations for the predictand weights are obtained as follows. First, from Eq. (17), $\mathbf{A}'\mathbf{A}'^{\mathrm{T}} = \mathbf{U}\mathbf{S}\mathbf{S}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}$, which means that $\mathbf{U}$ is the matrix of eigenvectors of $\mathbf{A}'\mathbf{A}'^{\mathrm{T}}$. The eigenvalues and eigenvectors of $\mathbf{A}'\mathbf{A}'^{\mathrm{T}}$ are found by solving the eigenvalue problem $\mathbf{A}'\mathbf{A}'^{\mathrm{T}}\,\mathbf{u} = s^2\mathbf{u}$, or in terms of the weight $\mathbf{q}_y = (\mathbf{Y}\mathbf{Y}^{\mathrm{T}})^{-1/2}\,\mathbf{u}$,

$$(\mathbf{Y}\mathbf{Y}^{\mathrm{T}})^{-1/2}\mathbf{A}'\mathbf{A}'^{\mathrm{T}}(\mathbf{Y}\mathbf{Y}^{\mathrm{T}})^{1/2}\mathbf{q}_y = s^2\mathbf{q}_y. \qquad \text{(A3)}$$

Then using the definition of $\mathbf{A}'$ in Eq. (13),

$$\begin{aligned}(\mathbf{Y}\mathbf{Y}^{\mathrm{T}})^{-1/2}\mathbf{A}'\mathbf{A}'^{\mathrm{T}}(\mathbf{Y}\mathbf{Y}^{\mathrm{T}})^{1/2} &= (\mathbf{Y}\mathbf{Y}^{\mathrm{T}})^{-1/2}\mathbf{Y}'\mathbf{X}'^{\mathrm{T}}\mathbf{X}'\mathbf{Y}'^{\mathrm{T}}(\mathbf{Y}\mathbf{Y}^{\mathrm{T}})^{1/2}\\ &= (\mathbf{Y}\mathbf{Y}^{\mathrm{T}})^{-1}\mathbf{Y}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\mathbf{Y}^{\mathrm{T}}. \qquad \text{(A4)}\end{aligned}$$

The eigenvalue problem in Eq. (A3) is

$$(\mathbf{Y}\mathbf{Y}^{\mathrm{T}})^{-1}\mathbf{Y}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\mathbf{Y}^{\mathrm{T}}\mathbf{q}_y = s^2\mathbf{q}_y, \qquad \text{(A5)}$$

which is Eq. (14.11) of von Storch and Zwiers (1999). The usual CCA equations for the predictor weights follow similarly from $\mathbf{A}'^{\mathrm{T}}\mathbf{A}' = \mathbf{V}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{V}^{\mathrm{T}}$.

REFERENCES

Akaike, H., 1973: Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory,* B. N. Petrov and F. Czaki, Eds., Akademiai Kiado, 267–281.

Boulesteix, A.-L., and K. Strimmer, 2007: Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief. Bioinform.,* **8,** 32–44.

Bretherton, C. S., C. Smith, and J. M. Wallace, 1992: An intercomparison of methods for finding coupled patterns in climate data. *J. Climate,* **5,** 541–560.

Browne, M. W., 2000: Cross-validation methods. *J. Math. Psychol.,* **44,** 108–132.

DelSole, T., and P. Chang, 2003: Predictable component analysis, canonical correlation analysis, and autoregressive models. *J. Atmos. Sci.,* **60,** 409–416.

——, and J. Shukla, 2006: Specification of wintertime North America surface temperature. *J. Climate,* **19,** 2691–2716.

——, and M. K. Tippett, 2007: Predictability: Recent insights from information theory. *Rev. Geophys.,* **45,** RG4002, doi:10.1029/2006RG000202.

Ehrendorfer, M., and J. Tribbia, 1997: Optimal prediction of forecast error covariances through singular vectors. *J. Atmos. Sci.,* **54,** 286–313.

Feddersen, H., A. Navarra, and M. N. Ward, 1999: Reduction of model systematic error by statistical correction for dynamical seasonal predictions. *J. Climate,* **12,** 1974–1989.

Glahn, H. R., 1968: Canonical correlation and its relationship to discriminant analysis and multiple regression. *J. Atmos. Sci.,* **25,** 23–31.

Golub, G. H., and C. F. Van Loan, 1996: *Matrix Computations.* 3rd ed. The Johns Hopkins University Press, 694 pp.

Gower, J. C., and G. B. Dijksterhuis, 2004: *Procrustes Problems.* Oxford University Press, 248 pp.

Hastie, T., A. Buja, and R. Tibshirani, 1995: Penalized discriminant analysis. *Ann. Stat.,* **23,** 73–102.

New, M. G., M. Hulme, and P. D. Jones, 2000: Representing twentieth-century space–time climate variability. Part II: Development of 1901–96 monthly grids of terrestrial surface climate. *J. Climate,* **13,** 2217–2238.

Roeckner, E., and Coauthors, 1996: The atmospheric general circulation model ECHAM-4: Model description and simulation of present-day climate. Tech. Rep. 218, Max-Planck Institute for Meteorology, Hamburg, Germany, 90 pp.

Schwarz, G., 1978: Estimating the dimension of a model. *Ann. Stat.,* **6,** 461–464.

Thacker, W. C., 1999: Principal predictors. *Int. J. Climatol.,* **19,** 821–834.

van den Dool, H., 2006: *Empirical Methods in Short-Term Climate Prediction.* Oxford University Press, 240 pp.

Vinod, H. D., 1976: Canonical ridge and econometrics of joint production. *J. Econometrics,* **4,** 147–166.

von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research.* Cambridge University Press, 494 pp.

Wang, X. L., and F. Zwiers, 2001: Using redundancy analysis to improve dynamical seasonal mean 500 hPa geopotential forecasts. *Int. J. Climatol.,* **21,** 637–654.

Widmann, M., 2005: One-dimensional CCA and SVD, and their relationship to regression maps. *J. Climate,* **18,** 2785–2792.

——, C. Bretherton, and E. P. Salathe Jr., 2003: Statistical precipitation downscaling over the northwestern United States using numerically simulated precipitation as a predictor. *J. Climate,* **16,** 799–816.

Wold, S., A. Ruhe, H. Wold, and W. J. Dunn III, 1984: The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM J. Sci. Stat. Comput.,* **5,** 735–743.

Yu, Z.-P., P.-S. Chu, and T. Schroeder, 1997: Predictive skills of seasonal to annual rainfall variations in the U.S. Affiliated Pacific Islands: Canonical correlation analysis and multivariate principal component regression approaches. *J. Climate,* **10,** 2586–2599.

Zucchini, W., 2000: An introduction to model selection. *J. Math. Psychol.,* **44,** 41–61.