Regression-based methods for finding coupled patterns

MICHAEL K. TIPPETT*
International Research Institute for Climate and Society, Palisades, NY, USA

TIMOTHY DELSOLE
George Mason University, Fairfax VA
Center for Ocean-Land-Atmosphere Studies, Calverton, MD, USA

SIMON J. MASON
International Research Institute for Climate and Society, Palisades, NY, USA

ANTHONY G. BARNSTON
International Research Institute for Climate and Society, Palisades, NY, USA

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 New York 10964, USA. (tippett@iri.columbia.edu)

ABSTRACT

There are a variety of multivariate statistical methods for analyzing the relations between two data sets. 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 patterns methods are equivalent to different diagonalizations of the regression between the two data sets. The different diagonalizations are computed using the singular value decomposition of the regression matrix developed using data that is 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 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 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 data set, finding the projections (empirical orthogonal functions; EOFs) or components that explain the most variance. Methods such as CCA and MCA work with two data sets, finding projections that optimize some form of linear association between the two data sets: CCA selects components of each data set so as to maximize their correlation; MCA does likewise, except to maximize their 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. 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 the predictands 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. Although the regression is invariant with respect to linear transformations of the data (as long as the predictor transformation is invertible), the SVD of the regression matrix is not invariant since linear

transformations change the definition of orthogonality. Therefore different transformations of the data yield distinct diagonalizations of the regression that diagnose particular properties of the regression. For instance, previous work has shown that when a whitening transformation is applied to the data, the SVD of the regression matrix corresponds to CCA (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 is suitably transformed for each method. The connection between the pattern methods and multivariate regression provides a common framework 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 the multivariate regression behaves 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 $\langle \cdot \rangle$ denotes expectation; we take x and y to be deviations from their respective means, and thus have zero-mean. The regression coefficient measures the strength of the relation between

predictor and predictand, and can be manipulated to obtain quantities such as correlation, explained variance and covariance. Specifically,

$$\begin{aligned} \text{correlation} &= \frac{\langle xy \rangle}{\sqrt{\langle x^2 \rangle \langle y^2 \rangle}} = a\sqrt{\frac{\langle y^2 \rangle}{\langle x^2 \rangle}}\,, \\ \sqrt{\text{explained variance}} &= \frac{\langle xy \rangle}{\sqrt{\langle x^2 \rangle}} = a\sqrt{\langle x^2 \rangle}\,, \\ \text{covariance} &= \langle yx \rangle = a\langle x^2 \rangle\,. \end{aligned} \tag{3}$$

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 also 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

$$\hat{y}' = a'x' = mal^{-1}x'\,; \tag{4}$$

the new regression coefficient a' is related to the original regression coefficient by $a' = mal^{-1}$. Combining Eqs. (3) and (4) 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 regression coefficient a' is the square-root of the variance explained by x;
- 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 interpretation of the regression coefficient, does not fundamentally change the regression. This concept is generalized to the multivariate case in section 3, and in section 4 we present the appropriate multivariate generalizations of these transformations of the data 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 $\mathbf{y}$ is linearly related to the multivariate predictor $\mathbf{x}$ by

$$\mathbf{y} = \mathcal{A}\mathbf{x} + \epsilon\,, \tag{5}$$

where $\mathcal{A}$ is a suitably dimensioned matrix and ϵ represents random noise which is independent of the predictor; $\mathbf{x}$ and $\mathbf{y}$ are anomaly fields, and we use the convention that $\mathbf{x}$ and $\mathbf{y}$ are column (rather than row) vectors. The least-squares estimate $\hat{\mathbf{y}}$ of the predictand is given by linear regression as

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

where the notation $()^T$ 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 i-th column is the i-th 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 i-th column is the i-th 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 $\mathbf{A} \equiv \left(\mathbf{Y}\mathbf{X}^T\right)\left(\mathbf{X}\mathbf{X}^T\right)^{-1}$.[1]

[1] We use the convention that the data in $\mathbf{X}$ and $\mathbf{Y}$ are normalized by the square-root of the number of samples. This convention simplifies the notation by making the sample averages have the same form as the expectations

As in the univariate case of Eq. (4), 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 variance is

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

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. (4). This relation provides several pieces of useful information. First, when the transformation 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. Re-scaling 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 greater than the number of samples, the sample covariance matrix $\mathbf{X}\mathbf{X}^T$ is singular, and reduction of the number of predictors is a way of regularizing the regression problem. The number of predictors is often reduced using principal component analysis (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 principal components (PCs) is called *pre-filtering* in the context of CCA (Bretherton et al. 1992). Remarkably, when $\mathbf{L}$ is the

with $\langle \mathbf{x}\mathbf{x}^T \rangle = \mathbf{X}\mathbf{X}^T$ and $\langle \mathbf{y}\mathbf{x}^T \rangle = \mathbf{Y}\mathbf{X}^T$.

pre-filtering transformation that maps the data onto a subset of its PCs, the regression developed with the pre-filtered data is the same as the original regression applied to the pre-filtered 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 $\mathbf{x}$ exceeds the number of samples and pre-filtering is done using the maximum number of PCs, the in-sample error is zero 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. Note that this procedure gives an overly optimistic estimate of skill due to selection bias (Zucchini 2000), and that 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 (and if necessary the number of predictands) has been reduced so that the number of predictors is less than the number of samples, and the predictor covariance is invertible.

Another important consequence of the relation in Eq. (8) follows from noting that the error variance $(\mathbf{y}' - \hat{\mathbf{y}}')^T(\mathbf{y}' - \hat{\mathbf{y}}')$ of the transformed variable is minimized, and that $(\mathbf{y}' - \hat{\mathbf{y}}')^T(\mathbf{y}' - \hat{\mathbf{y}}') = (\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{M}^T\mathbf{M})(\mathbf{y} - \hat{\mathbf{y}})$. Therefore, not only is the sum of squared error $(\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}})$ minimized, but so is the positive semi-definite quadratic function of the error $(\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{M}^T\mathbf{M})(\mathbf{y} - \hat{\mathbf{y}})$. Changing the weighting of the predictands does not change the least-squares estimate, and in fact, setting some of the weights to zero does not change the least-squares estimate since $\mathbf{M}$ is not required to be invertible. For instance, choosing $\mathbf{M} = e_i^T$

[2]There are in-sample estimates of the out-of-sample error such as AIC and BIC which take into account the number of predictors.

where e_i is the i-th column of the identity matrix means that the least-squares estimate minimizes $\|e_i^T(\mathbf{y} - \hat{\mathbf{y}})\|^2 = (y_i - \hat{y}_i)^2$ which is the error of the i-th component of the predictand. Therefore regression minimizes not only the total error variance but the error variance of each component separately. Consequently, the regression developed using all the components of the predictand simultaneously is the same as the regressions developed with individual components of the predictand separately.

This last property of regression aids the interpretation of principal component regression (PCR). In PCR, regressions are developed between predictor PCs and and each of the predictands individually. The above conclusion means that PCR is the same as developing the regression between the set of predictands and the PCs simultaneously. This shows a connection with canonical correlation analysis (CCA) since a CCA-based regression model with EOF pre-filtering 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 pre-filtering of the predictor and no other truncation such as pre-filtering 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. Then the SVD of the transformed regression matrix diagonalizes the regression and identifies projections of the data that maximize these measures. These projection 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:

$$\begin{aligned} \mathbf{x}' &= (\mathbf{X}\mathbf{X}^T)^{-1/2}\mathbf{x} \\ \mathbf{y}' &= (\mathbf{Y}\mathbf{Y}^T)^{-1/2}\mathbf{y}\,. \end{aligned} \tag{11}$$

The appearance of the inverse of the predictand covariance indicates that it may be necessary to pre-filter the predictand as well as the predictor. The matrix square-root is not uniquely defined; post-multiplication 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 components. Such a transformation is sometimes called a *whitening* transformation (DelSole and Tippett 2007) since the transformed data are uncorrelated and have unit variance

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

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

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

since $\mathbf{X}'\mathbf{X}'^T = \mathbf{I}$. The (i,j)-th element of $\mathbf{A}'$ is the correlation between the i-th component of $\mathbf{y}'$ and the j-th component of $\mathbf{x}'$, denoted y'_i and x'_j respectively, since

$$\mathbf{A}'_{ij} = e_i^T\mathbf{A}'e_j = e_i^T\mathbf{Y}'\mathbf{X}'^T e_j = e_i^T\mathbf{Y}'(e_j^T\mathbf{X}')^T = \langle y'_i x'_j\rangle\,, \tag{14}$$

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

Instead of looking at the correlations between individual components of $\mathbf{x}'$ and $\mathbf{y}'$ we can examine the correlation of 1-dimensional projections of the data. Projecting the transformed predictand and predictor data onto the vectors $\mathbf{u}$ and $\mathbf{v}$, respectively, gives the time-series

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

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

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

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

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 singular value decomposition (SVD) of $\mathbf{A}'$ is defined to be

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

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and $\mathbf{S}$ is a diagonal matrix with nonnegative entries s_i ordered from largest to smallest. The singular vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ are the i-th columns of $\mathbf{U}$ and $\mathbf{V}$ and satisfy

$$\mathbf{u}_i^T\mathbf{A}'\mathbf{v}_i = s_i\,. \tag{18}$$

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

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

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 projection vectors for the untransformed variables are the columns of the matrices $\mathbf{Q}_\mathbf{x}$ and $\mathbf{Q}_\mathbf{y}$ defined so that the projection of the untransformed variables is equal to the projection of the transformed variables

$$\mathbf{Q}_\mathbf{x}^T\mathbf{X} = \mathbf{V}^T\mathbf{X}'\ \text{and}\ \mathbf{Q}_\mathbf{y}^T\mathbf{Y} = \mathbf{U}^T\mathbf{Y}'\,. \tag{20}$$

Using Eq. (11) gives $\mathbf{Q}_\mathbf{y} = (\mathbf{Y}\mathbf{Y}^T)^{-1/2}\mathbf{U}$ and $\mathbf{Q}_\mathbf{x} = (\mathbf{X}\mathbf{X}^T)^{-1/2}\mathbf{V}$. Although the projection vectors for the transformed variables are orthogonal, the projection vectors for the untransformed

variables are not; or more precisely, they are orthogonal with respect to a different norm since $\mathbf{Q}_\mathbf{y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1}\mathbf{Q}_\mathbf{y} = \mathbf{I}$ and $\mathbf{Q}_\mathbf{x}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{Q}_\mathbf{x} = \mathbf{I}$. Consequently, the projection vectors are not the same as the patterns that multiply the corresponding time-series. Expressing the data as patterns $\mathbf{P}_\mathbf{x}$ and $\mathbf{P}_\mathbf{y}$ that multiply the time-series $\mathbf{Q}_\mathbf{x}^T\mathbf{X}$ and $\mathbf{Q}_\mathbf{y}^T\mathbf{Y}$, the patterns are found by solving

$$\mathbf{X} = \mathbf{P}_\mathbf{x}\mathbf{Q}_\mathbf{x}^T\mathbf{X} \quad \text{and} \quad \mathbf{Y} = \mathbf{P}_\mathbf{y}\mathbf{Q}_\mathbf{y}^T\mathbf{Y} \tag{21}$$

which gives $\mathbf{P}_\mathbf{x} = (\mathbf{X}\mathbf{X}^T)^{1/2}\mathbf{V}$ and $\mathbf{P}_\mathbf{y} = (\mathbf{Y}\mathbf{Y}^T)^{1/2}\mathbf{U}$. The pattern and projection vectors are orthogonal to each other.

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 $\mathbf{Q}_\mathbf{x}$ ($\mathbf{Q}_\mathbf{y}$) being the predictor (predictand) projection vectors, $\mathbf{P}_\mathbf{x}$ ($\mathbf{P}_\mathbf{y}$) the predictor (predictand) patterns, and the diagonal elements of $\mathbf{S}$ the canonical correlations (DelSole and Chang 2003, see Appendix of this paper for a derivation of the usual CCA equations). Using the relation between $\mathbf{A}$ and $\mathbf{A}'$ in (9), the regression matrix can be simply written using the projection vectors and patterns as

$$\mathbf{A} = (\mathbf{Y}\mathbf{Y}^T)^{1/2}\mathbf{A}'(\mathbf{X}\mathbf{X}^T)^{-1/2} = (\mathbf{Y}\mathbf{Y}^T)^{1/2}\mathbf{U}\mathbf{S}\mathbf{V}^T(\mathbf{X}\mathbf{X}^T)^{-1/2} = \mathbf{P}_\mathbf{y}\mathbf{S}\mathbf{Q}_\mathbf{x}^T\,. \tag{22}$$

The above relation shows that CCA diagonalizes the regression. Since $\mathbf{Q}_\mathbf{x}^T\mathbf{P}_\mathbf{x} = \mathbf{I}$, $\mathbf{A}\mathbf{P}_\mathbf{x} = \mathbf{P}_\mathbf{y}\mathbf{S}$, 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 $\mathbf{P}_\mathbf{y}$ and $\mathbf{Q}_\mathbf{x}$ 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 (Ehrendorfer and Tribbia 1997).

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 variance explained. The appropriate generalization to the multivariate problem is to apply the whitening

transformation to the predictor as in Eq. (11)

$$\mathbf{x}' = (\mathbf{X}\mathbf{X}^T)^{-1/2}\mathbf{x}\,, \tag{23}$$

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

$$\mathbf{A}' = \mathbf{Y}\mathbf{X}' \tag{24}$$

since $\mathbf{X}'\mathbf{X}'^T = \mathbf{I}$. Proceeding as in the previous section shows that that 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 $\mathbf{u}$ and $\mathbf{v}$ of the predictand and predictor data is

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

This ratio is maximized when $\mathbf{u}$ and $\mathbf{v}$ are respectively the leading left and right leading singular vectors of $\mathbf{A}'$. Therefore $s_1^2 = (\mathbf{u}_1^T\mathbf{A}'\mathbf{v}_1)^2$ is the maximum variance explained by a single predictor. Conversely, $\langle \mathbf{y}^T\mathbf{y}\rangle - s_1^2$ 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 $\langle \mathbf{y}^T\mathbf{y}\rangle - s_1^2 - s_2^2$. The variances add since the predictor projections are uncorrelated, a consequence of $\mathbf{v}_1^T\mathbf{X}'\mathbf{X}'^T\mathbf{v}_2 = \mathbf{v}_1^T\mathbf{v}_2 = 0$. The predictand projection time-series are correlated but the predictand projection (and pattern) vectors are orthogonal since $\mathbf{u}_1^T\mathbf{u}_2 = 0$. This decomposition of the data is called redundancy analysis (RDA; von Storch and Zwiers 1999; Wang and Zwiers 2001). Additional details of the projection 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.

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 find the projections that maximize the explained standardized variance. Explicitly, we use the transformations

$$\begin{aligned} \mathbf{x}' &= (\mathbf{X}\mathbf{X}^T)^{-1/2}\mathbf{x} \\ \mathbf{y}' &= (\mathrm{Diag}\,\mathbf{Y}\mathbf{Y}^T)^{-1/2}\mathbf{y}\,, \end{aligned} \tag{26}$$

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

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

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

$$\mathbf{A}'_{ij} = e_i^T\mathbf{A}'e_j = e_i^T\mathbf{Y}'\mathbf{X}'^T e_j = \frac{1}{\sqrt{e_i^T\mathbf{Y}\mathbf{Y}^T e_i}} e_i^T\mathbf{Y}(e_j^T\mathbf{X}')^T\,. \tag{28}$$

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 is

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

The $\mathbf{u}$ and $\mathbf{v}$ which 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 *principal predictor analysis* (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 projection 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 predictors by the inverse of their covariance:

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

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

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

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

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

This ratio is maximized when $\mathbf{u}$ and $\mathbf{v}$ are the left and right leading singular vectors of $\mathbf{A}'$. This decomposition of the data is maximum covariance analysis (MCA), sometimes referred to as SVD; we use the name MCA to distinguish between the regression scheme and the matrix decomposition (von Storch and Zwiers 1999).

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

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

where $\mathbf{Y}\mathbf{X}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T$ is a SVD.[4] To connect Eq. (33) with a commonly used linear model based on MCA modes (e.g., Widmann et al. 2003) we express the predictor data as $\mathbf{X} = \mathbf{V}\mathbf{B}$, 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 $\mathbf{B} = \mathbf{V}^T\mathbf{X}$. Substituting this representation of the predictor data into Eq. (33) gives

$$\mathbf{A} = \left(\mathbf{U}\mathbf{S}\mathbf{V}^T\right)\left(\mathbf{V}\mathbf{B}\mathbf{B}^T\mathbf{V}^T\right)^{-1} = \mathbf{U}\mathbf{S}\left(\mathbf{B}\mathbf{B}^T\right)^{-1}\mathbf{V}^T. \tag{34}$$

This form is similar to usual MCA-based linear models. However, usually $\mathbf{B}\mathbf{B}^T$ is replaced by its diagonal, the variance of the MCA time-series (e.g., Widmann et al. 2003). This approximation means that correlations between MCA modes are neglected, and the resulting estimate is not generally a least-squares estimate (LSE). The matrix $\mathbf{B}\mathbf{B}^T$ 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 $\mathbf{X}\mathbf{X}^T$ is proportional to the identity matrix; the latter condition is true when, for instance, the predictors are whitened variables. For this reason, Widmann (2005) 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. Therefore we call the method in Eq. (34) *LSE-MCA* since it uses the projections that maximize covariance like MCA but is a least-squares estimate. 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. Some implementations of *partial least-squares* (PLS) regression are the same as LSE-MCA (Boulesteix and Strimmer 2006) though earlier implementations (Wold et al. 1984) combine regression with an iterative SVD algorithm.

[4] Forming $\mathbf{Y}\mathbf{X}^T$ 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.

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. 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.

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 $\mathbf{A}' = \mathbf{U}\mathbf{S}\mathbf{V}^T$ is replaced with $\dot{\mathbf{A}}' = \mathbf{U}\dot{\mathbf{S}}\mathbf{V}^T$ where the first r diagonal elements of $\dot{\mathbf{S}}$ are the same as those of $\mathbf{S}$ and the rest are zero. The resulting regression 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 regres-

sion may optimize mutual information (CCA)[5], explained variance (RDA), the 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 2 components. Suppose that $\mathbf{Y}\mathbf{X}^T = \mathbf{I}$ and the MCA modes are the columns of the identity matrix. Then the regression matrix is simply

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

[5]This is a consequence of the facts that (i) mutual information of normally distributed 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.

The MCA approximation of the regression matrix is the diagonal matrix

$$\mathbf{A}_{\text{MCA}} = \text{Diag}(\mathbf{X}\mathbf{X}^T)^{-1} \,. \tag{36}$$

We take the predictor covariance to have the form

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

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, i.e., when $\theta = 0$. Additionally, suppose the predictand covariance has the similar structure

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

These choices for the covariances imply that the CCA modes are the same as the EOFs and that the canonical correlations are 0.8 and 0.5.

The error variance of the regression is

$$\begin{aligned} \langle \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 \rangle &= \text{tr}\left(\mathbf{A}\langle \mathbf{x}\mathbf{x}^T \rangle \mathbf{A}^T + \langle \mathbf{y}\mathbf{y}^T \rangle - \mathbf{A}\langle \mathbf{x}\mathbf{y}^T \rangle - \langle \mathbf{y}\mathbf{x}^T \rangle \mathbf{A}^T \right) \\ &= \text{tr}\left(\mathbf{A}\mathbf{X}\mathbf{X}^T\mathbf{A}^T + \mathbf{Y}\mathbf{Y}^T - \mathbf{A} - \mathbf{A}^T \right) \\ &= \text{tr}(\mathbf{Y}\mathbf{Y}^T - (\mathbf{X}\mathbf{X}^T)^{-1}) \\ &= \frac{1}{0.8^2} + \frac{1}{(0.5\sigma)^2} - 1 - \frac{1}{\sigma^2} \,, \end{aligned} \tag{39}$$

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

$$\begin{aligned} \langle \|\mathbf{A}_{\text{MCA}}\mathbf{x} - \mathbf{y}\|^2 \rangle &= \text{tr}\left(\mathbf{A}_{\text{MCA}}\mathbf{X}\mathbf{X}^T\mathbf{A}_{\text{MCA}}^T + \mathbf{Y}\mathbf{Y}^T - \mathbf{A}_{\text{MCA}} - \mathbf{A}_{\text{MCA}}^T \right) \\ &= \text{tr}(\mathbf{Y}\mathbf{Y}^T - (\mathbf{X}\mathbf{X}^T)^{-1}) + \text{tr}\left(\mathbf{A}_{\text{MCA}}\mathbf{X}\mathbf{X}^T\mathbf{A}_{\text{MCA}}^T - (\mathbf{X}\mathbf{X}^T)^{-1} \right) \\ &= \langle \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 \rangle + \text{tr}\left(\mathbf{A}_{\text{MCA}}\mathbf{X}\mathbf{X}^T\mathbf{A}_{\text{MCA}}^T - (\mathbf{X}\mathbf{X}^T)^{-1} \right) \,. \end{aligned} \tag{40}$$

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

$$\frac{\langle \|\mathbf{A}_{\text{MCA}}\mathbf{x} - \mathbf{y}\|^2 \rangle}{\langle \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 \rangle} = 1 + \frac{\text{tr}\left(\mathbf{A}_{\text{MCA}}\mathbf{X}\mathbf{X}^T\mathbf{A}_{\text{MCA}}^T - (\mathbf{X}\mathbf{X}^T)^{-1}\right)}{\langle \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 \rangle} \,. \tag{41}$$

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 $\mathbf{X}\mathbf{X}^T = \mathbf{I}$ and again the MCA modes are the same as the predictor EOFs. As σ decreases, the relative error of the MCA linear model increases.

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 2-dimensional, diagonal example where the correlations of the two components are specified and examine the error of rank-1 regressions as the variance of one of the components is varied. In particular, suppose that

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

The variance of the first and second components 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 $\mathbf{Y}\mathbf{X}^T$ is diagonal and given by

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

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

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

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

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

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 component 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 component 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 (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 that the squared error to be minimized by the rank-1 regression matrix accounting for the first component which has highest correlation. On the other hand, if σ is sufficiently large, then the rank-1 regression should be based on the second component.

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 component since it explains more covariance. However, the second component 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 2-dimensional example, the RDA rank-1 regression coincides with either the CCA or LSE-MCA rank-1 regressions, depending on which one has smaller 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 ECHAM 4.5 (Roeckner et al. 1996) T42 ensemble simulations of March-May (1950-2000) precipitation over the Northeast of Brazil. At this time of the year, precipitation over the Northeast of Brazil is 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 ($0.5^\circ \times 0.5^\circ$) rainfall observation data set (New et al. 2000). Leave-one-out cross-validation is used to select the level of EOF pre-filtering[6] as well as the number of patterns retained in the regression; the truncations for each method are chosen to maximize the sum over gridpoints 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).

[6]Predictor and predictand are pre-filtered in CCA. Only the predictor is pre-filtered in RDA and PPA.

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 which are captured by all the methods. The CCA regression uses 5 predictor EOFs and 2 predictand EOFs to form a rank-2 regression; RDA and PPA use rank-1 regressions based on 5 and 3 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 4 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 and the robustness of the methods is unknown. One could imagine poor estimation of the correlations among the predictors outweighing neglecting the true inter-predictor correlations. In any case, in this example, LSE-MCA does out-perform 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, and in fact is lower than that of the RDA regression.

8. Summary and conclusions

Two commonly used linear methods for finding coupled patterns in two data sets 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 variables and constructing linear prediction models. Pattern methods like redundancy analysis (RDA) and principal predictor analysis (PPA) were developed specifically for use in 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.

An important fact is that 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 semi-definite quadratic function of the error. This fact implies that developing the regressions with each predictand individually is the same as developing the regression with all the predictands simultaneously. Consequently, principal component regression (PCR) in which regressions are developed between predictor PCs and individual predictands is the same as the regression developed between the set of predictands and the predictor PCs simultaneously, which in turn is the same as CCA with EOF pre-filtering 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 interpretation 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 data is replaced by its normalized PCs plays an key role.

The structure of the regression matrix is revealed by the singular value decomposition (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 other than the usual one for the predictor and predictand (Ehrendorfer and Tribbia 1997).

Interestingly, we note that 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 and is equivalent to some implementations of partial least squares (Boulesteix and Strimmer 2006). There are some special cases when MCA and LSE-MCA are the same, such as when the MCA patterns are also EOFs of the predictor. In general, as illustrated in a 2-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 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 the worst performance of all the 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 2-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.

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 pre-filtering 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 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 number 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 sub-agencies.

APPENDIX

a. EOF pre-filtering

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

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

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

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

b. Alternative form for CCA

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

$$(\mathbf{Y}\mathbf{Y}^T)^{-1/2}\mathbf{A}'\mathbf{A}'^T(\mathbf{Y}\mathbf{Y}^T)^{1/2}\mathbf{q}_\mathbf{y} = s^2\mathbf{q}_\mathbf{y}\,. \tag{A3}$$

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

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

The eigenvalue problem is

$$(\mathbf{Y}\mathbf{Y}^T)^{-1}\mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{Y}^T\mathbf{q}_\mathbf{y} = s^2\mathbf{q}_\mathbf{y}\,, \tag{A5}$$

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

REFERENCES

Boulesteix, A., and K. Strimmer, 2006: Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief. Bioinform.*, **8**, 32–34.

Bretherton, C. S., C. Smith, and J. M. Wallance, 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.

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

DelSole, T., and M. K. Tippett, 2007: Predictability: Recent insights from information theory. *Rev. Geophys.*, in press.

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*. Third ed., The Johns Hopkins University Press, Baltimore, 694 pp.

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

New, M. G., M. Hulme, and P. D. Jones, 2000: Representing 20th century space-time climate variability. II: Development of 1901-1996 monthly terrestrial climate fields. *J. Climate*, **13**, 2217–2238.

Roeckner, E., K. Arpe, L. Bengtsson, M. Christoph, M. Claussen, L. Dümenil, M. Esch, M. Giorgetta, U. Schlese, and U. Schulzweida, 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.

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

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, West Nyack, NY, USA, 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.

Widmann, M., C. Bretherton, and E. P. Salathé, 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 Journal on Scientific and Statistical Computing*, **5**, 735–743.

Yu, Z.-P., P.-S. Chu, and T. Schroeder, 1997: Predictive skills of seasonal to annual rainfall

variations in the US 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.

List of Figures

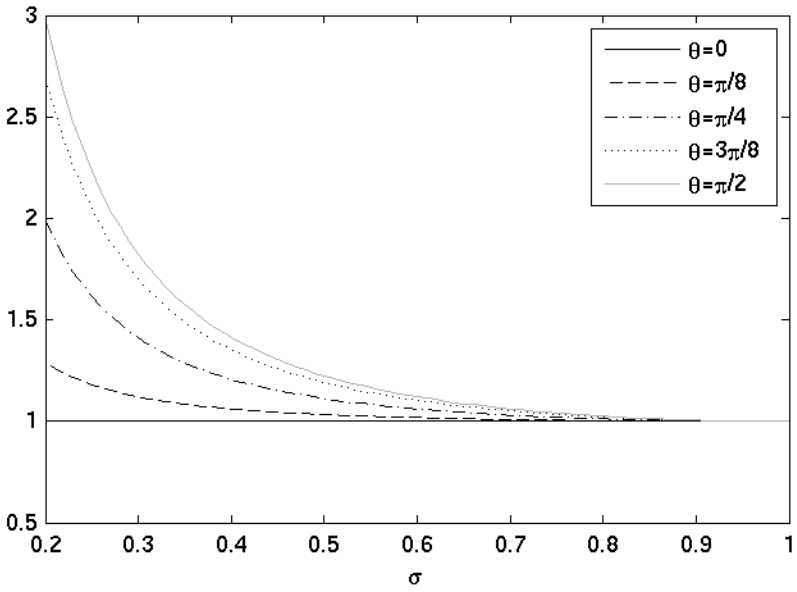

FIG. 1. Ratio of the MCA linear model error to that of the LSE-MCA regression as a function of σ for different values of the angle θ between predictor EOFs and MCA modes (see text).

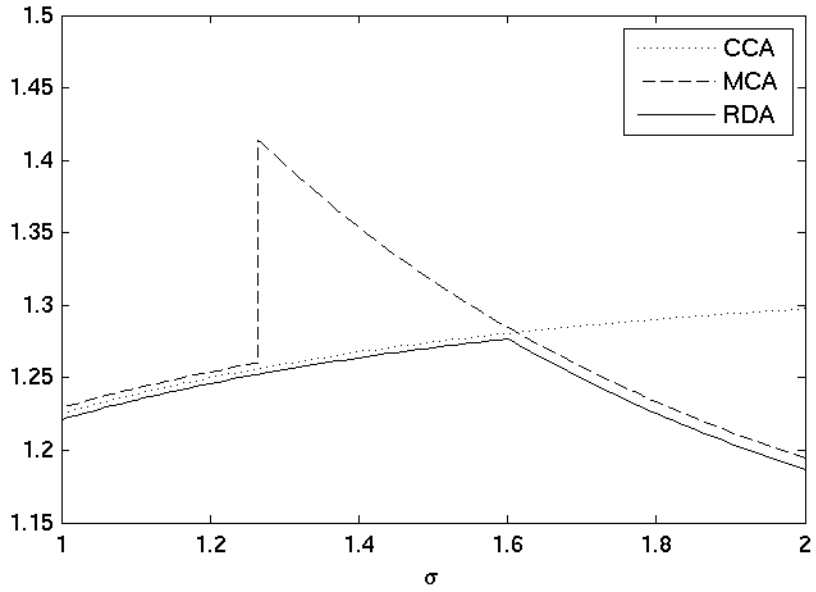

FIG. 2. Error of the rank-1 regression relative to that of the full regression as a function of σ for $c_1 = 0.8$ and $c_2 = 0.5$. Curves are offset for legibility.

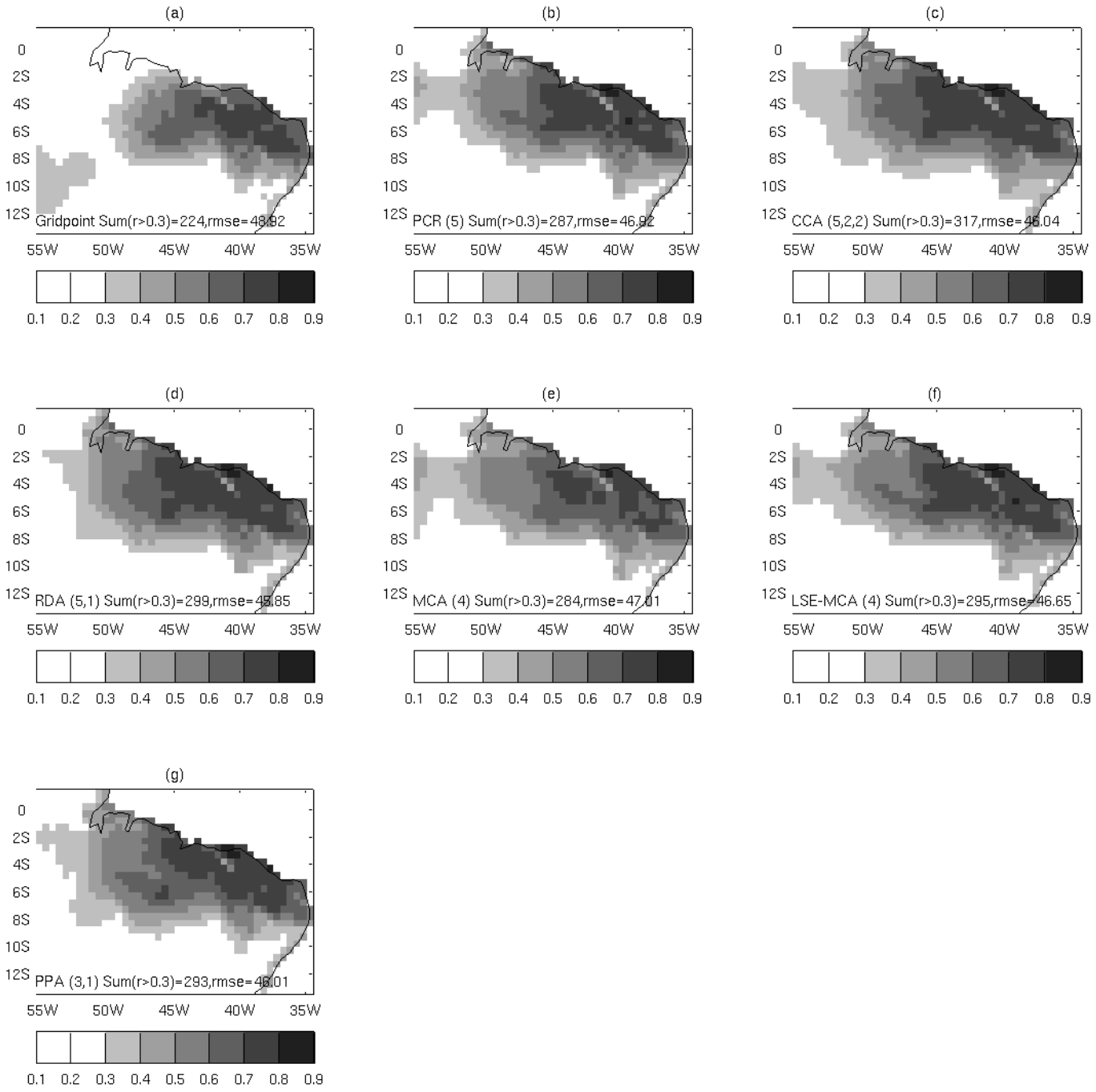

FIG. 3. Cross-validated correlation between corrected simulation and observed precipitation for (a) gridpoint regression, (b) PCR, (c) CCA, (d) RDA, (e) MCA, (f) LSE-MCA, and (g) PPA. Truncation (predictor EOFs, predictand EOFs, regression patterns), gridpoint sum of correlations greater than 0.3 and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

List of Tables

	CCA	RDA	PPA	MCA
optimizes	correlation	explained variance	sum of squared correlations	covariance
$\mathbf{x}'$	$\mathbf{x}' = (\mathbf{X}\mathbf{X}^T)^{-1/2}\mathbf{x}$	$\mathbf{x}' = (\mathbf{X}\mathbf{X}^T)^{-1/2}\mathbf{x}$	$\mathbf{x}' = (\mathbf{X}\mathbf{X}^T)^{-1/2}\mathbf{x}$	$\mathbf{x}' = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{x}$
$\mathbf{y}'$	$\mathbf{y}' = (\mathbf{Y}\mathbf{Y}^T)^{-1/2}\mathbf{y}$	$\mathbf{y}' = \mathbf{y}$	$\mathbf{y}' = (\mathrm{Diag}\,\mathbf{Y}\mathbf{Y}^T)^{-1/2}\mathbf{y}$	$\mathbf{y} = \mathbf{y}'$
projections	$\mathbf{Q}_\mathbf{x} = (\mathbf{X}\mathbf{X}^T)^{-1/2}\mathbf{V}$	$\mathbf{Q}_\mathbf{x} = (\mathbf{X}\mathbf{X}^T)^{-1/2}\mathbf{V}$	$\mathbf{Q}_\mathbf{x} = (\mathbf{X}\mathbf{X}^T)^{-1/2}\mathbf{V}$	$\mathbf{Q}_\mathbf{x} = (\mathbf{X}\mathbf{X}^T)^{-1/2}\mathbf{V}$
	$\mathbf{Q}_\mathbf{y} = (\mathbf{Y}\mathbf{Y}^T)^{-1/2}\mathbf{U}$	$\mathbf{Q}_\mathbf{y} = \mathbf{U}$	$\mathbf{Q}_\mathbf{y} = (\mathrm{Diag}\,\mathbf{Y}\mathbf{Y}^T)^{-1/2}\mathbf{U}$	$\mathbf{Q}_\mathbf{y} = \mathbf{U}$
patterns	$\mathbf{P}_\mathbf{x} = (\mathbf{X}\mathbf{X}^T)^{1/2}\mathbf{V}$	$\mathbf{P}_\mathbf{x} = (\mathbf{X}\mathbf{X}^T)^{1/2}\mathbf{V}$	$\mathbf{P}_\mathbf{x} = (\mathbf{X}\mathbf{X}^T)^{1/2}\mathbf{V}$	$\mathbf{P}_\mathbf{x} = (\mathbf{X}\mathbf{X}^T)^{1/2}\mathbf{V}$
	$\mathbf{P}_\mathbf{y} = (\mathbf{Y}\mathbf{Y}^T)^{1/2}\mathbf{U}$	$\mathbf{P}_\mathbf{y} = \mathbf{U}$	$\mathbf{P}_\mathbf{y} = (\mathrm{Diag}\,\mathbf{Y}\mathbf{Y}^T)^{1/2}\mathbf{U}$	$\mathbf{P}_\mathbf{y} = \mathbf{U}$

TABLE 1. The quantity optimized, the variable transformations, the projections, and the patterns for CCA, RDA, PPA, and MCA. In all cases, $\mathbf{U}\mathbf{S}\mathbf{V}^T$ is the SVD of the transformed regression $\mathbf{A}' = \mathbf{Y}'\mathbf{X}'^T$ and the decomposition of the original regression is $\mathbf{A} = \mathbf{P}_\mathbf{y}\mathbf{S}\mathbf{Q}_\mathbf{x}^T$.

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

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