AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE

10.1002/2013JD020382

Key Points:
- regPCovR is a new method for investigating the coupling between fields
- regPCovR balances the overprediction and underprediction found in other methods
- Coupling between the Australasian SLP field and rainfall field is investigated

Supporting Information:
- Figures S1 and S2
- Readme

Correspondence to:
M. J. Fischer,
mjf@ansto.gov.au

Citation:
Fischer, M. J. (2014), Regularized principal covariates regression and its application to finding coupled patterns in climate fields, *J. Geophys. Res. Atmos.*, *119*, 1266–1276, doi:10.1002/2013JD020382.

Received 18 JUN 2013
Accepted 15 JAN 2014
Accepted article online 17 JAN 2014
Published online 12 FEB 2014

Regularized principal covariates regression and its application to finding coupled patterns in climate fields

M. J. Fischer[1]

[1] Institute for Environmental Research, Australian Nuclear Science and Technology Organisation, Lucas Heights, New South Wales, Australia

Abstract There are many different methods for investigating the coupling between two climate fields, which are all based on the multivariate regression model. Each different method of solving the multivariate model has its own attractive characteristics, but often the suitability of a particular method for a particular problem is not clear. Continuum regression methods search the solution space between the conventional methods and thus can find regression model subspaces that mix the attractive characteristics of the end-member subspaces. Principal covariates regression is a continuum regression method that is easily applied to climate fields and makes use of two end-members: principal components regression and redundancy analysis. In this study, principal covariates regression is extended to additionally span a third end-member (partial least squares or maximum covariance analysis). The new method, regularized principal covariates regression, has several attractive features including the following: it easily applies to problems in which the response field has missing values or is temporally sparse, it explores a wide range of model spaces, and it seeks a model subspace that will, for a set number of components, have a predictive skill that is the same or better than conventional regression methods. The new method is illustrated by applying it to the problem of predicting the southern Australian winter rainfall anomaly field using the regional atmospheric pressure anomaly field. Regularized principal covariates regression identifies four major coupled patterns in these two fields. The two leading patterns, which explain over half the variance in the rainfall field, are related to the subtropical ridge and features of the zonally asymmetric circulation.

1. Introduction

Many areas of climate research involve the identification of coupled patterns in order to estimate the dependence between two climate fields. Examples include the following: statistical downscaling [*Li and Smith*, 2009], seasonal climate prediction [*Lim et al.*, 2011], climate field reconstruction using proxy records [*Smerdon et al.*, 2011], and linear inverse modeling and filtering [*Solomon and Newman*, 2012]. In estimating the coupling between two fields, there are several issues that arise:

1. The atmospheric state evolves with several degrees of freedom. This means that the variation in one field may be due to multiple patterns in another field. For example, precipitation over Europe is affected by multiple patterns in the atmospheric pressure field [*Qian et al.*, 2000].
2. A plethora of methods exist for estimating the coupling between fields, including principal components regression (PCR) [*Li and Smith*, 2009], redundancy analysis (RDA) [*Wang and Zwiers*, 2001], canonical correlation analysis (CCA) [*Tippett et al.*, 2008], partial least squares (PLS) [*Smoliak et al.*, 2010], and modern variants thereof [e.g., *Tenenhaus and Tenenhaus*, 2011; *Klami et al.*, 2013]. How to choose a method that is suitable for a particular problem is often unclear, and understanding the limitations and potential of each method with respect to different problems is an ongoing area of research.
3. The third issue is that climate field data can be incomplete, have zero values (e.g., precipitation), and the spatial dimension of the fields may be much larger than the number of temporal observations. This makes the estimation and statistical inference of climate field relationships more complicated [*DelSole and Yang*, 2011].

With respect to point 2, one possibility is to treat different methods as end-members along a continuum of regression solutions. Different field regression methods generally attempt to define a low-dimensional subspace of the predictor field, according to some criterion. The criterion could relate to variance in the predictor field, variance in the dependent (response) field, or to some other statistical property. To define a continuum of solutions, first, the dimension (d, the rank or number of components) of the subspace needs to be fixed. A continuum can then be achieved by weighting the solutions (of a given rank) for regression models that use distinct criterion. In this way, the regression models that use distinct criterion can be thought of as end-member solutions, between which there is a continuum of possible solutions. The predictive ability of solutions in this continuum can then be explored. In the statistical literature, regression models that explore such a range of solutions are known as continuum regression models.

The main aims of this study are to develop a new method for continuum regression that can be applied to climate fields (that may be incomplete or may have temporally sparse data) and to illustrate the method by example. The new method builds on an earlier method known as principal covariates regression (PCovR) [*de Jong and Kiers*, 1992]. PCovR is a continuum regression method that has two end-member solutions: PCR and RDA. The new method encompasses three end-members (PCR, RDA, and PLS) and can be used to efficiently investigate the predictability of regression models over a larger model space. The new method is illustrated by applying it to the problem of predicting a rainfall field from a local atmospheric pressure field [e.g., *Qian et al.*, 2000; *Li and Smith*, 2009; *Smoliak et al.*, 2010], focusing on the southern Australian region.

2. Methods

2.1. The General Model

The general linear model for predicting field Y from field X is as follows:

$$Y = XW\,P_Y + E \tag{1}$$

where X is an $n \times p$ predictor field, Y is an $n \times q$ response field, W is a $p \times d$ matrix of weights, P_Y is a $d \times q$ matrix of regression weights, and E is an $n \times q$ matrix of residuals (see also Table 1). The residuals are generally assumed to be from a multivariate normal distribution that is independent in time and space. Note that X and Y are typically anomaly fields (see section 3 for more information). The central idea is to project X into a low-dimensional space or subspace (T, an $n \times d$ matrix) that retains useful information for predicting Y (i.e., $T = XW$ and $Y = TP_Y$). Once T is determined using a particular method, P_Y can always be estimated by least squares principles. The matrix W has the general form:

$$W = AV\,L^{-0.5} \tag{2}$$

where A is a transformation matrix and V and L are the eigenvectors and diagonalized eigenvalues of the covariance matrix of XA. The transformation matrix could take many forms, e.g., $A = I_p$, $A = S_{XX}^{-1}S_{XY}$, or even $A = \left[c_1 I_p,\ c_2 S_{XX}^{-1}S_{XY}\right]$. Note I_p is an identity matrix of rank p, S_{XX} is a cross-product matrix of the subscripted variables, and c_i represents scalar weightings.

The choice of A corresponds to different regression methods. This section briefly reviews how A and T are defined for several methods, starting with PCR and RDA. From this, it is apparent that RDA can be extended by adding a simple parameter, which allows the investigation of a continuum of predictor subspaces, including a PLS-type subspace. A third method, principal covariates regression, is also discussed and is extended here in a similar way to RDA in order to investigate an even wider range of subspaces.

The focus here is on PCR, RDA, and PLS because these methods are robust against missing values in Y. This robustness occurs because the estimated subspace lies within the column space of X. No subspace of Y is directly required as in some other methods, e.g., conventional canonical correlation analysis. Hence, the latter method is not used in this study.

2.2. Principal Components Regression

In PCR, $A = I_p$, so the subspace T is estimated straightforwardly from the singular value decomposition of the predictor field X:

$$X = TL^{0.5}V' \tag{3}$$

where T and V are both orthonormal matrices. Thus, the leading components of T maximally account for the variation in X:

$$R_X^2 = tr(X'P_T X)tr(S_{XX})^{-1} \tag{4}$$

but do not necessarily explain much variance in another field, Y.

Table 1. Notation of Variables

Notation	Definition
X	$n \times p$ predictor field, p stations with n time points
Y	$n \times q$ response field, q stations with n time points
E	$n \times q$ matrix of residuals, i.e., $E = Y - \hat{Y}$
T	$n \times d$ matrix consisting of d component time series
W	$p \times d$ matrix of weights (or loadings), i.e., $T = XW$
A	$p \times k$ transformation matrix, k may vary in size
V	$k \times d$ matrix of eigenvectors
L	Diagonal matrix of eigenvalues, i.e., $L = \text{diag}(\lambda)$
P_X	$d \times p$ matrix of regression weights, i.e., $X = TP_X$
P_Y	$d \times q$ matrix of regression weights, i.e., $Y = TP_Y$
P_T	$n \times n$ projection matrix, i.e., $P_T = TS_{TT}^{-1}T'$
S_{XX}	Cross-product matrix of the subscripted variables
I_p	Identity matrix of rank p
$tr()$	Matrix trace, e.g., $tr(S_{XX}) = \|X\|^2$
$\hat{e}(Y\|X)$	The residuals from the regression of Y on X
$\text{Cor}(Y\|X)$	Scalar correlation = $tr(S_{YX}S_{XX}^{-1}S_{XY})tr(S_{YY})^{-1}$

In PCR, a step-function filter is typically applied to the components:

$$T = XW\text{diag}(f), \quad \text{where} \tag{5a}$$

$$f_i = \begin{cases} 1 & i \leq d \\ 0 \end{cases} \tag{5b}$$

and d is the number of components retained. Thus, components that are associated with small singular values are truncated.

2.3. Redundancy Analysis and Regularization

In RDA, $A = S_{XX}^{-1}S_{XY}$, which is a transformation matrix that maximizes the squared correlation between XA and Y. The estimation of T can be performed in two steps. First, the predicted field ($\hat{Y}$) is calculated using XA. Then, T is estimated from the singular value decomposition of $\hat{Y}$:

$$XS_{XX}^{-1}S_{XY} = TL^{0.5}V'. \tag{6}$$

Like the principal components, the RDA components (T) exist in the column space of X, but unlike the principal components of X, the RDA components maximally account for the variation in Y:

$$R_Y^2 = tr(Y'P_TY)tr(S_{YY})^{-1}. \tag{7}$$

Note that the leading RDA components do not necessarily explain much of the variance in X.

A major issue with RDA occurs when $p > n$, because then S_{XX}^{-1} (in A) cannot be estimated consistently. One solution to this problem is to estimate the predicted field ($\hat{Y}$) using $X(S_{XX} + kI_p)^{-1}S_{XY}$, which is known as ridge regression [*Jones*, 1972]. The effect of the ridge parameter k is to reduce the magnitude of regression coefficients that are associated with principal components of X characterized by small eigenvalues. The implementation of RDA with ridge regression can be termed ridge-RDA.

The value of ridge-RDA can be understood by considering the correlation between the ridge-RDA components and Y:

$$R_Y^2 = \text{Cor}(Y|X(S_{XX} + kI_p)^{-1}S_{XY}V). \tag{8}$$

When $k \to 0$, ridge-RDA finds the components that maximize the correlation between the components and Y, as in conventional RDA. When $k \to \infty$, the above correlation becomes $\text{Cor}(Y|XS_{XY}V)$, because $(S_{XX} + kI_p)^{-1}$ approximates a scalar matrix. Note that $\text{Cor}(Y|XS_{XY}V) < \text{Cor}(Y|X(S_{XX})^{-1}S_{XY}V)$, because the correlation between the components and Y is optimal only for the conventional RDA solution ($k = 0$). Also, for $k \to \infty$, the ridge-RDA solution is similar to implementing partial least squares by singular value decomposition (PLS-SVD) [*Bougeard et al.*, 2008]. When $0 < k < \infty$, ridge-RDA provides a continuum of solutions which give different weighting to explained variance in the predictor and response fields.

2.4. Principal Covariates Regression and Regularization

Before discussing principal covariates regression, it is useful to consider PCR and RDA when the size (or rank) of the low-dimensional subspace is fixed. Ideally, we would like to find the simplest subspace of X that usefully predicts Y. For a subspace of a given rank, $\text{Cor}(Y|T_{\text{RDA}}) \geq \text{Cor}(Y|T_{\text{PCR}})$, while $\text{Cor}(X|T_{\text{PCR}}) \geq \text{Cor}(X|T_{\text{RDA}})$. Thus, it seems that T_{RDA} may provide the simplest predictive subspace in a computationally efficient manner, while PCR may underpredict the response field for a solution of a given rank. The main issue with PCR is that there is no efficient way of finding a subspace, of any rank, that maximizes $\text{Cor}(Y|T_{\text{PCR}})$, because we do not know a priori which combination of principal components of X to select. But RDA is also not without problems. Since RDA maximizes the correlation between T and Y, an RDA model that is calibrated using only one set of training data will most certainly overpredict the response field for the training period. Thus, for a solution of a given rank, it seems that PCR may tend to underpredict Y, while RDA may tend to overpredict Y. But between these two solutions, we may find subspaces that balance the overprediction and underprediction for a given rank.

Principal covariates regression is a method that combines PCR and conventional RDA [*de Jong and Kiers*, 1992]. The first step, as in RDA, is to estimate the predicted field $\hat{Y}$. Next, the predicted field is concatenated with the predictor field (forming Z), and Z is factored:

$$Z = \left[a^{0.5}/\|X\|, (1-a)^{0.5}\hat{Y}/\|\hat{Y}\|\right] \tag{9a}$$

$$Z = TL^{0.5}V' \tag{9b}$$

where a is a scalar weighting ($0 < a < 1$). Like T_{RDA} and T_{PCR}, the leading components of Z (T_{PCovR}) also exist in the column space of X, because $\hat{Y}$ (in Z) consists of linear combinations of the X columns. Hence, the leading components of Z will maximize:

$$aR_X^2 + (1-a)R_Y^2. \tag{10}$$

As $a \rightarrow 1$ the components will be more similar to the principal components of X, while as $a \rightarrow 0$ the components will be more similar to RDA components. Between the extremes are perhaps the more interesting cases, because then the components T will summarize X well and predict Y well.

Like conventional RDA, the above method of implementing PCovR is hindered when $p > n$, because estimating $\hat{Y}$ (in Z) requires S_{XX}^{-1}. A simple solution is to estimate $\hat{Y}$ using ridge regression. This method is different from conventional PCovR, because now the leading components of Z will optimize:

$$aR_X^2 + (1-a)\mathrm{Cor}(Y|X(S_{XX} + kI_p)^{-1}S_{XY}V). \tag{11}$$

For a fixed number of components $T_{1:d}$, this optimization now involves two parameters: a and k. Note that henceforth T refers to T_{regPCovR} (unless otherwise stated), and d is the number of components of T. The parameter a controls how well the components T explain variance in Y or X, reflecting a ridge-RDA or PCR solution at the extremes ($0 < a < 1$). The parameter k also controls how well the components T explain variance in Y (reflecting an RDA solution when $a, k = 0$), but unlike a the parameter k aims to retain components that are both stable and explain some variance in Y ($k \rightarrow \infty$, reflecting a PLS-type solution). Thus, regularized PCovR (regPCovR) encompasses PCR, RDA, and PLS.

The following section explains how the parameters a, k, and d were estimated in the present study.

2.5. Estimating the regPCovR Parameters *a*, *k*, and *d*

2.5.1. Cross Validation to Estimate *a* and *k*

The parameters a and k were estimated using cross validation. The procedure was performed using training and test sets which were generated by leaving out one season (e.g., leaving out the June-July-August (JJA) data in 1 year) and repeating for each study year (e.g., 1960–2002). Each predicted test set ($\hat{Y}_{\mathrm{test}}$) was then combined to formed the overall prediction $\hat{Y}_{\mathrm{TEST}}$:

$$\hat{Y}_{\mathrm{test}} = (X_{\mathrm{test}} - \bar{X}_{\mathrm{train}})B_{\mathrm{train}} + \bar{Y}_{\mathrm{train}} \tag{12}$$

$$\mathrm{vec}(Y) = \mathrm{diag}(\hat{Y}_{\mathrm{TEST}})b + e \tag{13a}$$

$$\mathrm{diag}(\hat{Y}_{\mathrm{TEST}}) = \begin{bmatrix} y_1 & 0 & 0 & 0 \\ 0 & y_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & y_q \end{bmatrix} \tag{13b}$$

where $B = WP_Y$, $\mathrm{vec}(Y)$ is an (nq) vector with the columns of Y stacked on top of each other, and $\mathrm{diag}(\hat{Y}_{\mathrm{TEST}})$ is an $(nq) \times q$ matrix with each column of $\hat{Y}_{\mathrm{TEST}}$ (i.e., $y_1, y_2, \ldots, y_q$) positioned along the diagonal.

The cross-validation statistic employed was the R^2 value (i.e., the multiple correlation coefficient) of the regression model in equation (13a). This statistic is termed R_{CV}^2. Although other metrics could be chosen, R^2 is a useful cross-validation metric because it relates to the information content between fields, whereas other

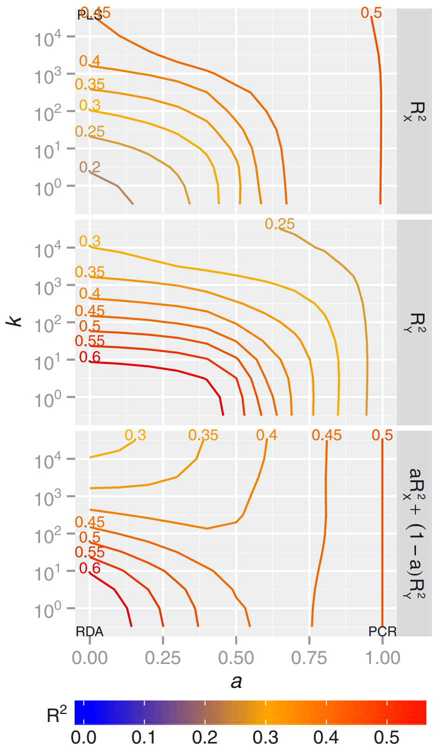

Figure 1. The relationships between the subspaces of X (estimated with regPCovR), and the X and Y fields (where X = Australasian SLP$_{JJA}$ field and Y = southern Australian rainfall P$_{JJA}$ field). (top and middle) The correlation between a rank 2 subspace $T_{(a,k)}$ and the X and Y fields, respectively (equations (4) and (7) in text) are shown. For each plot, the right edge of the diagram corresponds to a PCR subspace, and the bottom left and top left corners correspond to an RDA and PLS subspace, respectively (these labels have been added to the outer plots only). Hence, Figures 1 (top) and 1 (middle) show that the PCR subspace fits X well, but not Y. The RDA subspace fits Y well, but not X. (bottom) As one moves away from the RDA solution in the direction of the PLS or PCR solution, there are subspaces that can fit both X and Y reasonably (equations (10) and (11)). In all plots, the contour interval is 0.05 units.

metrics (e.g., mean squared error (MSE)) relate to the squared distance between fields. Metrics like MSE are thus affected by the variance of $\hat{Y}$.

The cross validation was performed by using a grid search over the sets $a \in \{0, 0.1, \dots, 1\}$ and $k = 10^j$ where $j \in \{-0.5, 0, 0.5, \dots, 4.5\}$. Also, for the cross validation, the number of regPCovR components was fixed at $d = 2$.

Note that two-deep cross validation [e.g., *DelSole*, 2007] is not used in this study, because it is computationally expensive when there is more than one parameter for selection by cross validation. The advantage of two-deep cross validation is that it provides a better estimate of out-of-sample predictive skill and information about the variance of the estimated parameters. In this study, the cross-validation skill estimated may be biased relative to the true out-of-sample predictive skill. Methods for speeding up two-deep cross validation for multiple parameter problems are currently being investigated.

2.5.2. A Partial Regression Method to Estimate *d*

Although d could be included as a free parameter in the cross validation, a computationally more efficient solution is to make d a fixed parameter for the cross validation, and a free parameter later. If d is initially fixed at a low value (e.g., $d = 1$ or 2), then we can test if there is any advantage in making d larger, by employing a partial regression approach.

Given a multiple predictor regression model such as $y = b_0 + X_{1:m}b_{1:m} + X_{m+1}b_{m+1}$, a conventional partial regression plot is a plot of the residuals of the regression of y on $X_{1:m}$ versus the residuals of $X_{m+1}|X_{1:m}$. (Note that the regression of y on X can be generically expressed as $y|X$, and the regression residuals as $\hat{e}(y|X)$.) In a partial regression plot, the slope of the least squares fitted line matches the regression coefficient b_{m+1} and the Pearson correlation coefficient calculated from the point coordinates is the partial correlation coefficient for the variable X_{m+1} in the multiple regression model. Thus, the partial regression plot indicates whether adding X_{m+1} to the regression significantly improves the regression or not.

The ideas of partial regression can be applied to the problem of estimating the number of significant regPCovR components ($T_{1:d}$) in the general linear model (equation (1)). That is, can the regression model be improved by adding an extra dimension to the subspace? An issue here is that in the general linear model, Y is multivariate, so it is difficult to plot the residual matrix of $Y|T_{1:m}$ on the y axis. A solution to this problem is to replace the residual matrix with its first principal component:

$$\hat{e}(Y|T_{1:m})v_1 = \hat{e}(T_{m+1}|T_{1:m})b_{m+1} + e \qquad (14)$$

where v_1 is the first empirical orthogonal function (EOF) (or the first column of the loading matrix) for the eigenvalue decomposition of $\hat{e}(Y|T_{1:m})$. Methods for the eigenvalue decomposition of matrices with missing values can be found in *Josse and Husson* [2012], if required. By examining plots of equation (14) for different values of m (starting at $m = 1$), the dimension of subspace T can be chosen: i.e., when b_{m+1} is not significant, then $d = m$. This method is computationally efficient and is applied to real data in the following section.

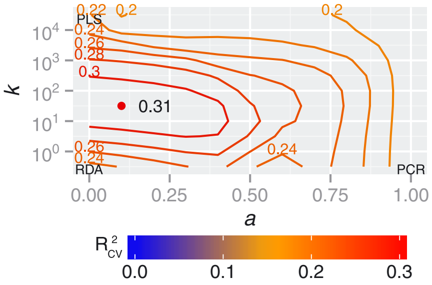

Figure 2. The cross-validation score for different regPCovR models (equation (13a)). Models with a higher cross-validation score have better out-of-sample prediction (although this may be biased relative to true out-of-sample prediction). The PCR, RDA, and PLS subspaces are found on the respective side or corners of the plot. The contour interval is 0.02 units.

3. An Application of regPCovR

3.1. Data

This application of regPCovR explores the coupling between two fields: the Australasian mean sea level pressure (SLP) field and the southern Australian rainfall (P) field, for the austral winter months (JJA) and for the years 1960–2002. These data and years were chosen for compatibility with a future study, which will involve other data sets (e.g., rainfall isotopes) and will be reported elsewhere.

Monthly mean sea level pressure data were extracted from the HADSLP2 data set with a horizontal resolution of 5 × 5° [*Allen and Ansell*, 2006], for the region 90 to 200°E and 60 to 10°S. The total number of grid points in this region is 220. These data contain no missing values.

Monthly land rainfall amount data were extracted from the Global Precipitation Climatology Centre's Full Data reanalysis version 6 with a horizontal resolution of 2.5 × 2.5° [*Becker et al.*, 2013], for the region 115 to 145°E and 32.5 to 45°S. The total number of grid points is 18 (see Figure S1 in the supporting information). There are no missing values, but over the years 1960–2002, there are 99 months with zero rainfall (i.e., 99 out of a possible $9288 = 12$ months $\times 43$ years $\times 18$ grid points). Prior to detrending and deseasonalizing these data, these 99 months were set as "missing" (i.e., given a "not available" value), so all the following analysis is being done on months with nonzero rainfall. Note that of the 99 months with missing values, only four were found in the austral winter (JJA) months.

Both data sets were detrended and deseasonalized using the method described in Appendix A. This is important because the focus here is on the coupling of the anomaly fields of the predictor and response variables. As the raw data contain both seasonal variance and trend variance, performing the analysis on the raw data would mean that the extracted relationships between the fields would carry a mixed signal (that due to trend + seasonal variance + anomaly) rather than a pure signal (that due to the anomalies only).

After the data sets were detrended and deseasonalized, the June, July, and August anomalies were extracted to form the austral winter anomaly fields (X and Y). The columns of X and Y were scaled to unit variance and area-weighted by the square root of cos(latitude) [*Baldwin et al.*, 2009].

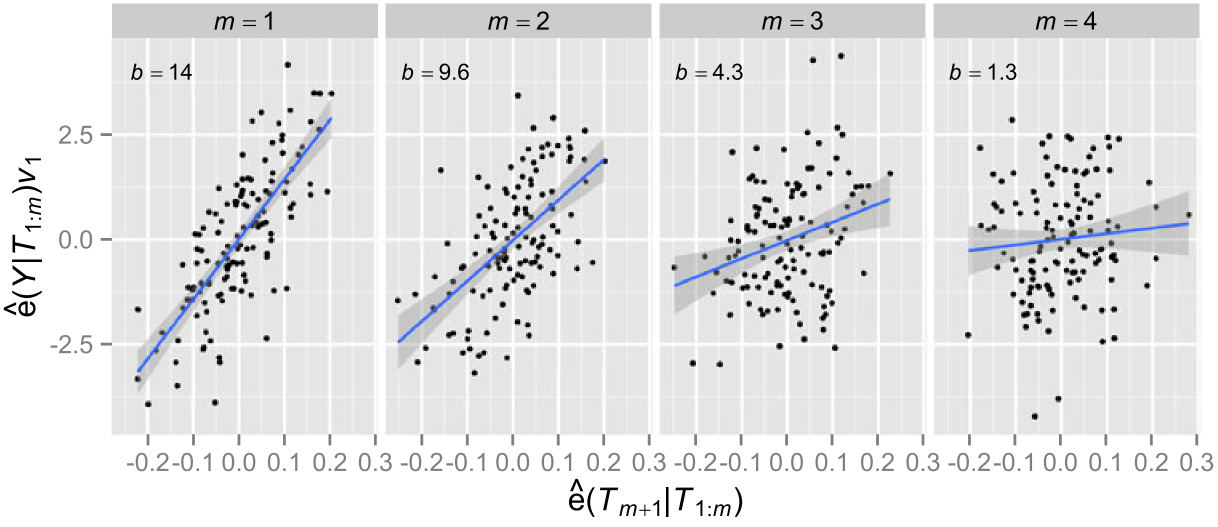

Figure 3. A partial regression plot for a multivariate regression model. The plot shows whether increasing the rank of the regression model from m to $m + 1$ significantly improves the regression model. The plot achieves this by exploring whether there is a significant relationship between the first principal component of the residual matrix for a rank m model (on the y axis), with the $m + 1$ subspace (equation (14)). In each plot, b is the slope of the partial regression line.

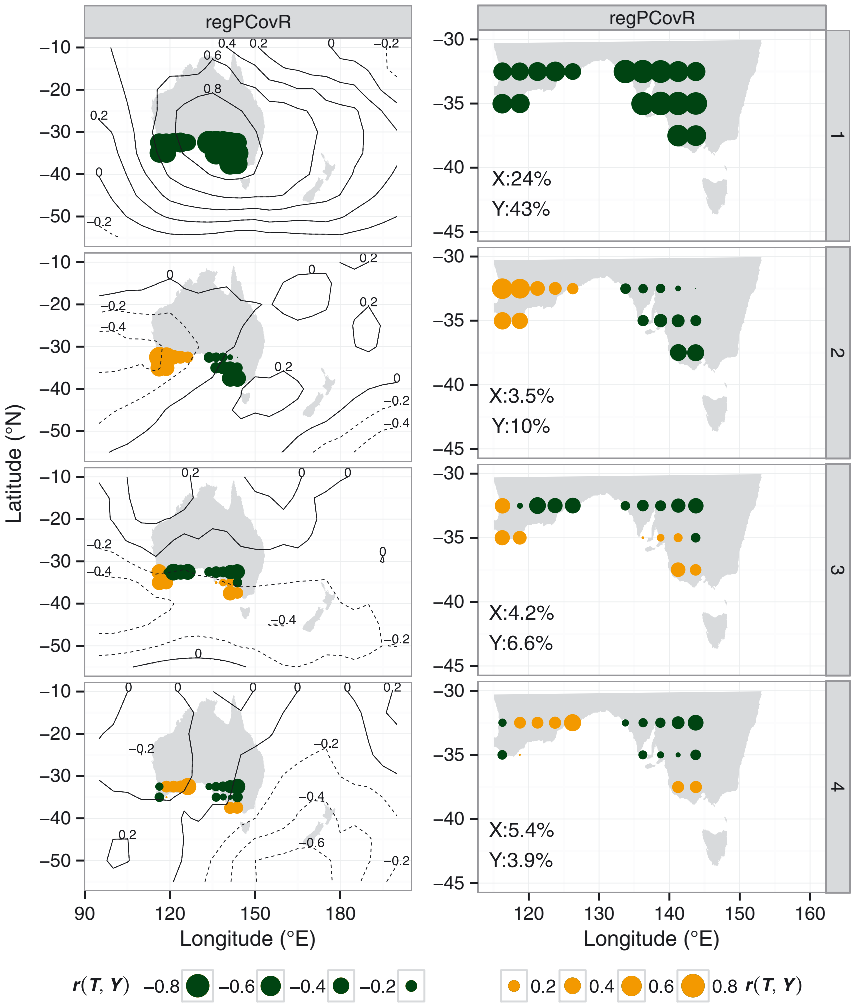

Figure 4. The first four leading components of Z (i.e., the subspace $T_{1:4}$), for the regPCovR model with the highest cross-validation score. (left) The contours show the correlation between the components of T and the gridded SLP$_{\mathrm{JJA}}$ anomaly field (contour interval is 0.2 correlation units, using Pearson's correlation). Positive and negative SLP anomalies are shown by solid and dashed contours. (right) The map zooms in on southern Australia, so that the precipitation response (to the corresponding SLP patterns) is more visible. The filled circles show the correlation between the components of T and the gridded southern Australian rainfall (P$_{\mathrm{JJA}}$) anomaly field. Positive and negative rainfall anomalies are shown by orange and green circles; the circle areas are proportional to the size of the correlation with T. The percentages (shown on the plots in Figure 4, right) refer to the proportion of variance that the respective component explains in the X and Y fields.

3.2. Australasian SLP and Southern Australian Rainfall

RegPCovR was used to investigate the relationship between the Australasian SLP$_{\mathrm{JJA}}$ field and southern Australian P$_{\mathrm{JJA}}$ field. In order to explore the properties of different regPCovR models, the subspace $T_{1:2}$ was estimated over the grid of a, k values (section 2.5.1), using the full X and Y fields for model calibration (equation (9a)). The resulting values of R^2_X, R^2_Y, and $aR^2_X + (1-a)R^2_Y$ are shown in Figure 1 and can be used to characterize the spectrum of regPCovR solutions. Models with high R^2_Y but low R^2_X are found at the grid point $a = 0, k = 0$ (RDA). Models with high R^2_X but low R^2_Y are found where $a = 1$ (PCR). Models which explain more equitable variance in both X and Y are found where $a = 0$ and $k > 10^3$ (PLS). Since this analysis is based only on within-sample correlations (i.e., using the full fields for model calibration), it does not show which model achieves the best cross-validation skill.

Using the cross-validation procedure in section 2.5.1, the model with the highest R^2_{CV} was found at the point $a = 0.1, k = 10^{1.5}$ (Figure 2). This shows that in this example, the rank 2 subspaces estimated from PCR, RDA, and PLS (on the corners or sides of Figure 2) all have a lower R^2_{CV} than other rank 2 solutions in the regPCovR solution space. The optimal solution, though, does lie near the ridge-RDA path ($a = 0, 0 < k < \infty$). This analysis also demonstrates that the two SLP patterns that best explain (predict) southern Australian winter precipitation are not found among the leading principal components of the SLP field.

Next, the partial regression method (section 2.5.2) was used to investigate if adding extra dimensions to the subspace significantly improved the regression model. Figure 3 shows the plots of $\hat{e}(Y|T_{1:m})v_1$ against $\hat{e}(T_{m+1}|T_{1:m})$ for four different values of m. As expected, the slope of this partial regression (b on the plots) becomes weaker as m increases, because each successive dimension explains a smaller proportion of the variance in Y. Therefore, the optimal number of components of T is four, because adding a fifth dimension to T, as in the last plot, does not significantly improve the regression model (the line $y = 0$ falls within the partial regression confidence limits).

The leading four coupled patterns for the SLP and precipitation anomaly fields are shown in Figure 4. The two leading components of T explain over half the variance in the precipitation field, but only 27.5% of the variance in the SLP field. The first coupled pattern consists of a broad high-pressure anomaly over Australia associated with a negative rainfall anomaly across the south. In the opposite phase, a low-pressure anomaly would be associated with positive rainfall anomalies. The second coupled pattern consists of a zonally asymmetric pressure anomaly (low-pressure anomaly in the west and a high-pressure anomaly east of Tasmania) associated with positive rainfall anomalies in the southwest but not southeast. The third coupled pattern consists of a meridional low-pressure anomaly (40–45°S) associated with positive rainfall anomalies on the southern fringe, but not in the southern interior. The fourth coupled pattern consists of a low-pressure anomaly in the southwest Pacific associated with a mainly positive rainfall anomaly in southwest Australia and fringe southeast, but not interior southeast and far southwest. These coupled patterns will be explored in more detail in the following section.

4. Discussion

Studies investigating the relationship between a predictor field and a response field, typically either (i) extract a single "climate" index (e.g., Southern Oscillation Index) from the predictor field using a limited number of stations, and compare this with the response field, or (ii) adopt a multivariate approach focused on one or more end-member subspaces (e.g., principal components regression or partial least squares). The regPCovR method is different from these approaches because it uses the full predictor and response fields (cf. case (i) above), and it aims to find the subspace which best captures the between-field relationships by exploring a range of model subspaces, not simply end-member subspaces (cf. case (ii)).

Several persistent features have been identified in the Southern Hemisphere atmosphere, such as the tropical oscillation (i.e., the Southern Oscillation), the subtropical ridge, and the subpolar zonal waves [e.g., *Ropelewski and Jones*, 1987; *Larsen and Nicholls*, 2009; *Hobbs and Raphael*, 2010]. In order to compare the regPCovR components with these features of the Southern Hemisphere circulation, indices of these features will be projected into the regPCovR subspace (this projection is described below). First, various Southern Hemisphere Climate Indices (SHCIs) are discussed.

In this study the following Southern Hemisphere climate indices were considered: Southern Oscillation Index (SOI) [*Ropelewski and Jones*, 1987], Southern Annular Mode (SAM) [*Marshall*, 2003; *Ho et al.*, 2012], Trans Polar Index (TPI) [*Jones et al.*, 1999], Subtropical Ridge (STR) [*Larsen and Nicholls*, 2009], Zonal Waves 1 and 3 (ZW1 and ZW3) [*Raphael*, 2004; *Hobbs and Raphael*, 2007], and the Southwest and Southeast Pacific Anticyclones (SWPA and SEPA) [*Hobbs and Raphael*, 2010].

These indices all describe persistent features of the Southern Hemisphere atmosphere. The source and/or calculation of these indices are detailed in the supporting information. Some indices were obtained from electronic archives, but indices for several features (ZW1, ZW3, SWPA, and SEPA) were recalculated from the gridded HADSLP2 Southern Hemisphere pressure field [*Allen and Ansell*, 2006]. The recalculation was done using more objective and efficient methods than in the studies cited above (see supporting information). Note that the indices for some features (SOI, SAM, and TPI) are expressed as univariate time series, but the indices for all the other features are multivariate time series. Specifically, ZW1, ZW3, SWPA, and SEPA are each described by three time series, which mark the longitude (or phase), latitude, and amplitude of the feature. Whether a feature is expressed as a univariate time series or as multivariate time series depends on whether the feature is considered to be stationary or quasi-stationary. For each feature, the austral winter anomaly amplitude time series was correlated against the Southern Hemisphere winter SLP anomaly field,

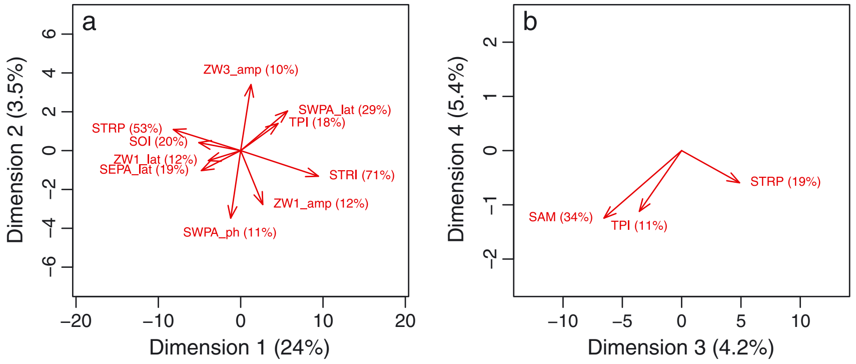

Figure 5. Biplots showing the correlation between the components of T and various indices of Southern Hemisphere climate variability (SHCIs, the vectors). The vectors (arrows) were calculated by projecting each SHCI into the subspace formed by the regPCovR components. (a) The winter SLP subspace formed by components T_1 and T_2. (b) The winter SLP subspace formed by components T_3 and T_4. In both panels, the percentage for each SHCI (on the arrows) refers to the variance explained by regressing each SHCI on the respective regPCovR components (e.g., equation (15)). The percentage for each dimension (on the plot axes) refers to the proportion of variance in the winter SLP field that each regPCovR component explains.

and the correlation maps are plotted in Figure S2 of the supporting information. The correlation maps show the expected dipole (SAM, ZW1, TPI, and SOI) or tripole (ZW3) patterns, and specific features (STR, SWPA, and SEPA).

Figure 5 shows the projections of the SHCIs into the regPCovR subspace. Prior to projection, the SHCI time series were deseasonalized and the seasonal anomalies for months JJA were extracted and scaled to unit variance. The $\mathrm{SHCI}_{\mathrm{JJA}}$ anomaly time series were projected into a regPCovR subspace by regressing them against two regPCovR components (which make up the axial dimensions of the subspace). For example, in Figure 5a, the projection of STRI (the intensity of the STR) into the subspace formed from the two leading regPCovR components gives the equation:

$$\mathrm{STRI} = 9.4T_1 + -1.32T_2, r^2 = 71\%. \tag{15}$$

The regression coefficients represent the projection (so the STRI vector points into the bottom right quadrant in Figure 5a), and the r^2 value is a measure of how close a particular SHCI lies to the subspace into which it is projected (vectors that have a low r^2 value are pointing away from the planes shown in Figure 5). Also, the correlation coefficient between any two indices shown on the diagram is given by the cosine of the angle between the corresponding vectors or axes. Vectors separated by small angles or opposite angles are positively and negatively correlated, while vectors that are perpendicular are independent. Note though, the scale of the horizontal and vertical axes in Figure 5 is not the same (so the diagram does not become squashed), and hence, there may be some visual distortion.

The first regPCovR component T_1 (or dimension 1) is strongly associated with the intensity and position of the subtropical ridge (STRI and STRP). Note that the position and intensity of the ridge are anticorrelated: a stronger STR anomaly occurs when its position is further southward than normal [*Larsen and Nicholls*, 2009]. The second regPCovR component T_2 (dimension 2) is partly associated with the zonally asymmetric components of the Southern Hemisphere circulation, especially the position (latitude and longitude) of the southwest Pacific anticyclone (SWPA_lat and SWPA_ph). Thus, here the leading two regPCovR components are not associated with the Southern Annular Mode (the SAM is the first EOF of the Southern Hemisphere extratropical SLP anomaly field). The third regPCovR component T_3 is related to a contrast between the amplitude of the SAM and the position of the subtropical ridge (STRP). That is, more rain tends to fall on the southern Australian fringe when the SAM is in a negative phase and the position of the STR is further northward. This type of superposition effect is not apparent in studies which consider only one phenomena/index. The fourth regPCovR component T_4 is not closely related to any of the above atmospheric indices, although it explains 5.4% of the variance in the predictor field. It appears similar to PC3 of the predictor field (Z3 in *Li and Smith* [2009]), with a low-pressure anomaly southwest of New Zealand, associated with a positive rainfall anomaly on the southeast fringe, but a negative rainfall anomaly in far southwest Australia.

The interannual variability of winter rainfall in southern Australia has been previously related to the SAM [*Nicholls*, 2010], zonal wave 3 [*van Ommen and Morgan*, 2010], and the subtropical ridge [*Larsen and Nicholls*, 2009]. These studies considered only within-sample correlations, hence, the out-of-sample predictive skill of the climate phenomena-rainfall relationships described by those studies is unclear. In contrast, *Li and Smith* [2009] used cross validation to determine the optimal number of components in a PCR model using a similar pressure field to our study and more localized rainfall fields. (Note that Li and Smith used the National Centers for Environmental Prediction-National Center for Atmospheric Research sea level pressure field and Australian Bureau of Meteorology rainfall field, whereas we have used HADSLP2 and GPCC rainfall.) The leading two PCR components (in *Li and Smith* [2009]) explained 65% of the variance in the predictor field, compared to the leading two regPCovR components which explain 27.5% of the variance in the predictor field (Figure 4). For a given number of components, the regPCovR model would tend to explain more of the variance in the response field than in the predictor field (Figures 1, 2, and 4); hence, the regPCovR model is likely to require fewer components than a PCR model to explain the same amount of variance in the response field. This is because the cross-validation step in regPCovR searches for a subspace, rather than the number of components in a predefined subspace (such as in PCR), which maximizes predictive skill.

In conclusion, regularized principal covariates regression is a continuum regression method that can be used to investigate the coupling between two fields. It has several attractive features, including that: it easily applies to response fields that have missing values or are temporally small, it explores a range of model spaces (so one does not have to choose a particular regression method), and it seeks a model subspace that will typically have a better predictive ability for a given number of components compared to other regression methods. In future work, the method will be expanded in order to investigate complex signal fields [*Schreier*, 2008], cross-validation parameter variance, and the influence of particular time points on the model subspace [*Barrett*, 2003]. Regularized principal covariates regression should be a useful tool in many areas of climate research that involve the coupling between fields.

Appendix A: Detrending and Deseasonalizing the Fields

A climate field can be thought of as being composed of several components:

$$\text{Field} = \text{Seasonal} + \text{Trend} + \text{Mean} + \text{Anomaly} + \text{noise}. \tag{A1}$$

The Mean component is the temporal mean of the field, the Trend and Seasonal components represent long-term and cyclical changes in the mean field, and the Anomaly component contains interannual variance. In order to examine the Anomaly field, the Trend and Seasonal components need to be modeled and removed. In this study, the Trend component is modeled as a linear function of time, while the Seasonal component is modeled using harmonic functions. This can be described by the following equation:

$$\text{Field} = \left[\sin(w_1 t), \cos(w_1 t), \sin(2w_1 t), \cos(2w_1 t)\right] B_2 + tB_1 + B_0 + \text{Anomaly} \tag{A2}$$

where Field is a matrix with time points as rows and grid points (or stations) as columns, t is a time vector, i.e., $t = \text{year}+\text{month}/12-1/24$, $w_1 = 2\pi\text{radians yr}^{-1}$ is an angular frequency, B_2 is a $4\times p$ matrix ($p =$ number of stations) containing the coefficients of the four harmonic functions in equation (A2), and B_1 and B_0 are row vectors (of dimension $1 \times p$) which contain the linear trend coefficients and the mean field, respectively.

By using the model in equation (A2), the Seasonal and Trend components can be simultaneously estimated. Equation (A2) is solved using the standard methods of multivariate linear regression. The Seasonal, Trend, and Mean components can then be removed from the climate field, leaving a field that is an Anomaly field (which also includes noise).

References

Allen, R. J., and T. Ansell (2006), A new globally complete monthly historical gridded mean sea level pressure data set (HadSLP2): 1850–2004, *J. Clim.*, *19*, 5816–5842.

Baldwin, M. P., D. B. Stephenson, and I. T. Joliffe (2009), Spatial weighting and iterative projection methods for EOFs, *J. Clim.*, *22*, 234–243, doi:10.1175/2008JCLI2147.1.

Barrett, B. E. (2003), Understanding influence in multivariate regression, *Commun. Stat. Theory Methods*, *32*, 667–680, doi:10.1081/STA-120018557.

Becker, A., P. Finger, A. Meyer-Christoffer, B. Rudolf, K. Schamm, U. Schneider, and M. Ziese (2013), A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901–present, *Earth Syst. Sci. Data*, *5*, 71–99, doi:10.5194/essd-5-71-2013.

Bougeard, S., M. Hanafi, and E. M. Qannari (2008), Continuum redundancy-PLS regression: A simple continuum approach, *Comput. Stat. Data Anal.*, *52*, 3686–3696, doi:10.1016/j.csda.2007.12.007.

de Jong, S., and H. A. L. Kiers (1992), Principal covariates regression: Part I. Theory, *Chemom. Intell. Lab. Syst.*, *14*, 155–164, doi:10.1016/0169-7439(92)80100-i.

DelSole, T. (2007), A Bayesian framework for multimodel regression, *J. Clim.*, *20*, 2810–2826, doi:10.1175/JCLI4179.1.

DelSole, T., and X. Yang (2011), Field significance of regression patterns, *J. Clim.*, *24*, 5094–5107, doi:10.1175/2011JCLI4105.1.

Ho, M., A. S. Kiem, and D. C. Verdon-Kidd (2012), The Southern Annular Mode: A comparison of indices, *Hydrol. Earth Syst. Sci.*, *16*, 967–982, doi:10.5194/hess-16-967-2012.

Hobbs, W. R., and M. N. Raphael (2007), A representative time-series for the Southern Hemisphere zonal wave 1, *Geophys. Res. Lett.*, *34*, L05702, doi:10.1029/2006GL028740.

Hobbs, W. R., and M. N. Raphael (2010), Characterizing the zonally asymmetric component of the SH circulation, *Clim. Dyn.*, *35*, 859–873, doi:10.1007/s00382-009-0663-z.

Jones, T. A. (1972), Multiple regression with correlated independent variables, *Math. Geol.*, *4*, 203–218.

Jones, P. D., M. J. Salinger, and A. B. Mullan (1999), Extratropical circulation indices in the Southern Hemisphere based on station data, *Int. J. Climatol.*, *17*, 1301–1317.

Josse, J., and F. Husson (2012), Handling missing values in exploratory multivariate data analysis methods, *J. Soc. Fr. Stat.*, *153*, 79–99.

Klami, A., S. Virtanen, and S. Kaski (2013), Bayesian canonical correlation analysis, *J. Mach. Learn. Res.*, *14*, 899–937.

Larsen, S. H., and N. Nicholls (2009), Southern Australian rainfall and the subtropical ridge: Variations, interrelationships and trends, *Geophys. Res. Lett.*, *36*, L08708, doi:10.1029/2009GL037786.

Li, Y., and I. Smith (2009), A statistical downscaling model for southern Australia winter rainfall, *J. Clim.*, *22*, 1142–1158, doi:10.1175/2008JCLI2160.1.

Lim, Y., S. Jo, J. Lee, H.-S. Oh, and H.-S. Kung (2011), An improvement of seasonal climate prediction by regularized canonical correlation analysis, *Int. J. Climatol.*, *32*, 1503–1512, doi:10.1002/joc.2368.

Marshall, G. J. (2003), Trends in the Southern Annular Mode from observations and reanalyses, *J. Clim.*, *16*, 4134–4143, doi:10.1175/1520-0442(2003)016<4134:TITSAM>2.0.CO;2.

Nicholls, N. (2010), Local and remote causes of the southern Australian autumn-winter rainfall decline, 1958–2007, *Clim. Dyn.*, *34*, 835–845, doi:10.1007/s00382-009-0527-6.

Qian, B., J. Corte-Real, and H. Xu (2000), Is the North Atlantic Oscillation the most important atmospheric pattern for precipitation in Europe?, *J. Geophys. Res.*, *105*(D9), 11,901–11,910, doi:10.1029/2000JD900102.

Raphael, M. N. (2004), A zonal wave 3 index for the Southern Hemisphere, *Geophys. Res. Lett.*, *31*, L23212, doi:10.1029/2004GL020365.

Ropelewski, C. F., and P. D. Jones (1987), An extension of the Tahiti-Darwin Southern Oscillation Index, *Mon. Weather Rev.*, *115*, 2161–2165, doi:10.1175/1520-0493(1987)115<2161:AEOTTS>2.0.CO;2.

Schreier, P. J. (2008), A unifying discussion of correlation analysis for complex random vectors, *IEEE Trans. Signal Process.*, *56*, 1327–1336, doi:10.1109/TSP.2007.909054.

Smerdon, J. E., A. Kaplan, E. Zorita, J. F. Gonzalez-Rouco, and M. N. Evans (2011), Spatial performance of four climate field reconstruction methods targeting the Common Era, *Geophys. Res. Lett.*, *38*, L11705, doi:10.1029/2011GL047372.

Smoliak, B. V., J. M. Wallace, M. T. Stoelinga, and T. P. Mitchell (2010), Application of partial least squares regression to the diagnosis of year-to-year variations in Pacific Northwest snowpack and Atlantic hurricanes, *J. Geophys. Res.*, *37*, L03801, doi:10.1029/2009GL041478.

Solomon, A., and M. Newman (2012), Reconciling disparate twentieth-century Indo-Pacific ocean temperature trends in the instrumental record, *Nat. Clim. Change*, *2*, 691–699, doi:10.1038/nclimate1591.

Tenenhaus, A., and M. Tenenhaus (2011), Regularized generalized canonical correlation analysis, *Psychometrika*, *76*, 257–284, doi:10.1007/s11336-011-9206-8.

Tippett, M. K., T. DelSole, S. J. Mason, and A. G. Barnston (2008), Regression-based methods for finding coupled patterns, *J. Clim.*, *21*, 4384–4398, doi:10.1175/2008JCLI2150.1.

van Ommen, T. D., and V. Morgan (2010), Snowfall increase in coastal East Antarctica linked with southwest Western Australian drought, *Nat. Geosci.*, *3*, 267–272, doi:10.1038/ngeo761.

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