Comparison of Probabilistic Statistical Forecast and Trend Adjustment Methods for North American Seasonal Temperatures

DANIEL S. WILKS

Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York

(Manuscript received 20 September 2013, in final form 18 December 2013)

ABSTRACT

The three multivariate statistical methods of canonical correlation analysis, maximum covariance analysis, and redundancy analysis are compared with respect to their probabilistic accuracy for seasonal forecasts of gridded North American temperatures. Derivation of forecast error covariance matrices for the methods allows a probabilistic formulation for the forecasts, assuming Gaussian predictive distributions. The three methods perform similarly with respect to probabilistic forecast accuracy as reflected by the ranked probability score, although maximum covariance analysis may be preferred because of its slightly better forecast skill and calibration. In each case the forecast accuracy for North American seasonal temperatures compares favorably to results from previously published studies. In addition, two alternative approaches are compared for alleviating the cold biases in the forecasts that derive from ongoing climate warming. Adding lagging 15-yr means to forecast temperature anomalies improved forecast accuracy and reduced the cold bias in the forecasts, relative to using the more conventional lagging 30-yr mean.

1. Introduction

Seasonal forecasts are often made with linear multivariate statistical methods. Although the dynamics of the climate system are nonlinear, in practice it has been found that nonlinear statistical methods perform no better than the traditional linear statistics for seasonal prediction (Tang et al. 2000; van den Dool 2007). Possibly this phenomenon occurs because the time averaging inherent in long-lead forecasting renders both the predictors and predictands Gaussian or quasi Gaussian under the central limit theorem (e.g., Johnson and Wichern 2002; Mardia et al. 1979), which in turn induces linear or quasi-linear relationships among them (Yuval and Hsieh 2002; Hsieh 2009).

Statistical seasonal forecasts are most frequently based on methods that seek to estimate joint patterns of variation between predictor and predictand fields, such as canonical correlation analysis (CCA; e.g., Barnett and Preisendorfer 1987; Barnston 1994; Wilks 2008) and maximum covariance analysis (MCA; e.g., Rodwell and Folland 2002; Uvo et al. 1998; Wilks 2008). As noted by von Storch and Zwiers (1999) both CCA and MCA are “symmetric” in their treatment of the predictors and predictands, in the sense that there is no distinction between them in the computations. Accordingly, the joint patterns that are estimated may not be optimal for forecasting, where the representation of the predictand field is paramount and fidelity of representation of the predictor field is not specifically of interest. They suggest that an alternative method, redundancy analysis (RA; Tyler 1982; van den Wollenberg 1977), could be particularly useful in forecast settings because it is designed to maximize predictand variance. Although RA has been used in the climate literature for statistical downscaling (WASA Group 1998; Pfizenmayer and von Storch 2001), postprocessing dynamical seasonal forecasts (Wang and Zwiers 2001), climate-field reconstruction (Kauker et al. 2008), and climate diagnostics (Bakalian et al. 2010), its performance in a seasonal-forecast setting has apparently not been investigated to date. In a similar way, Tippett et al. (2008) compare the mathematical properties of CCA, MCA, and RA in the absence of sampling noise and compare their performance in a downscaling example, apparently the methods have not yet been compared in the context of seasonal forecasting.

An important element of any uncertain forecast is explicit expression of that uncertainty, particularly as...
the additional information content of probability forecasts renders them more valuable to their users (e.g., Thompson 1962; Murphy 1977; Krzysztofowicz 1983). Probabilistic formats are especially warranted in inherently low-skill settings such as seasonal forecasts (e.g., Mason and Mimmack 2002). Nearly all previous forecasts based on CCA and MCA have been nonprobabilistic, however, with the predictand specified in terms of its forecast mean (vector) only. A CCA formulation in which both the mean vector and the forecast covariance matrix are computed has been described recently by Wilks (2014), who noted that the prediction covariance matrix specification could be extended also to other linear multivariate forecasting procedures such as MCA. Together the mean vector and covariance matrix allow specification of full multivariate Gaussian probability distributions for the predictand anomalies. The central limit theorem supports the assumption of multivariate Gaussian predictand distributions for seasonal averages, at least for “well behaved” variables such as temperature.

Another issue that must be faced when formulating seasonal forecasts in the present environment is the nonstationarity of the underlying climatic mean. Recent increases in seasonal temperature means have been large enough relative to the predictive resolution of seasonal forecasts that much of the realized seasonal forecast skill (relative to an aging climatological baseline) for the United States has derived from this warming trend (Livezey and Timofeyeva 2008; Peng et al. 2012), and neglecting or underrepresenting this trend has led to substantial cold forecast biases (Barnston et al. 2010; Barnston and Mason 2011; Goddard et al. 2003; Livezey and Timofeyeva 2008; Peng et al. 2012, 2013; Wilks 2000, 2008, 2013b; Wilks and Godfrey 2002). Adding a simple linear trend predictor to conventional CCA and MCA seasonal forecast formulations (Wilks 2008) decreased but did not eliminate this cold bias, because most of the recent climate warming has occurred since the mid-1970s and therefore has been faster in recent years than overall during historical training periods (Livezey et al. 2007; Wilks and Livezey 2013). A variety of alternative climate “normal” formulations have been proposed and evaluated for their skill in projecting recent warming into the near future (Livezey et al. 2007; Wilks 2013a; Wilks and Livezey 2013). The best current single method for representing nonstationarity in U.S. seasonal temperature means is the most recent 15-yr average, although these projections lag the ongoing warming and more sophisticated approaches may become preferable as the warming continues to unfold (Wilks 2013a; Wilks and Livezey 2013). Annually updated alternative climate normals such as most recent 15-yr averages can be incorporated into probabilistic seasonal forecasting either by simply adding them to the means of forecast distributions for the climate anomalies (Higgins et al. 2004), or by adjusting forecast probabilities through a recently proposed “naïve Bayesian” approach (Krakauer et al. 2013).

This paper will develop procedures for MCA- and RA-based probability forecasts, following the approach taken in Wilks (2014) for CCA forecasts, and will compare the three methods for producing probabilistic seasonal forecasts of gridded North American temperature averages using Indo-Pacific sea surface temperatures (SSTs) as predictors. The forecast procedures are trained using predictand anomalies relative to their most recent 30-yr averages. The benefits of also extrapolating recent warming trends captured by the most recent 15-yr averages are examined, using both simple additive adjustments to the means of the Gaussian forecast distributions and the naïve Bayesian approach. The mathematical development for the three forecasting methods is contained in section 2. The data sources and experimental protocols are described in section 3. The results are presented in section 4, and section 5 contains conclusions.

2. Forecast methods

a. Joint covariance matrix for the predictor and predictand principal components

The linear statistical forecast procedures employed here all involve manipulations of the joint sample covariance matrix for the [leading principal components (PCs) of the] gridded predictor and predictand variables,

\[
S = \langle (x_1, x_2, \ldots, x_I, y_1, y_2, \ldots, y_J)(x_1, x_2, \ldots, x_I, y_1, y_2, \ldots, y_J)^T \rangle
= \begin{pmatrix}
S_{xx} & S_{xy} \\
S_{yx} & S_{yy}
\end{pmatrix}.
\]

Here, the \((I + J)\)-dimensional column vector \((x_1, x_2, \ldots, x_I, y_1, y_2, \ldots, y_J)^T\) is a concatenation of the leading \(I\) centered predictor PCs \(x_i\) and the leading \(J\) centered predictand PCs \(y_j\), the angle brackets indicate an average taken over \(n\) training years, and the superscript \(T\) denotes vector or matrix transpose. To be specific, the leading \(I\) predictor PCs for year \(t\) are computed as

\[
x_t = E_t^I P_t^I,
\]

where the \(I\) columns of the matrix \(E\) contain the leading predictor eigenvectors and the vector \(P_t^I\) contains the
gridded SST predictor anomalies relative to 1951–80 for a given season in year \( t \). The leading \( J \) predictand PCs for year \( t \) are

\[
y_i = E_i^T T_i',
\]

where the \( J \) columns of the matrix \( E_j \) contain the leading \( J \) predictand temperature eigenvectors. The vector of gridded predictand temperature anomalies \( T_i' \) is centered using lagging 30-yr averages,

\[
T_i' = T_i - T_i^{(30)} = T_i - \frac{1}{30} \sum_{r=1}^{30} T_r,
\]

so that the predictand values for a target year are not used to center the forecast anomalies pertaining to themselves.

As indicated in Eq. (1), the joint covariance matrix \( S \) can be partitioned into the \((I \times I)\) covariance submatrix for the predictors \( S_{xx} \), the \((J \times J)\) covariance submatrix for the predictands \( S_{yy} \), and the \((I \times J)\) submatrix \( S_{xy} = S_{yx} \) of covariances between predictor and predictand elements. The constraint \( I \geq J \) is enforced in order that the matrix \( B \) (as defined for each forecast method in section 2b) onto which the predictand PCs are projected will be square and therefore invertible, allowing resynthesis of the dimensional predictand values.

b. Probabilistic multivariate statistical forecast methods

Each of the three forecast methods used here operates by projecting the leading predictor and predictand PCs \( x \) and \( y \) onto patterns chosen to optimize a particular criterion, yielding the vectors of linear combinations

\[
v = A^T x
\]

and

\[
w = B^T y
\]

for the predictors and predictands, respectively. The \( j \)th column of the \((I \times J)\) matrix \( A \), representing the \( j \)th predictor pattern onto which the predictor PCs are projected, relates to the \( j \)th element of the predictor linear combination vector \( v \). In a similar way, the \( j \)th column of the \((J \times J)\) matrix \( B \), representing the \( j \)th predictand pattern onto which the predictand PCs are projected, relates to the \( j \)th element of the predictand linear combination vector \( w \). The three forecasting methods differ in the ways that the pattern matrices \( A \) and \( B \) are computed.

At the forecasting stage the elements of the predictand linear combinations \( w \) are predicted from the elements of \( v \), using regressions. Having computed the quantities in Eq. (5) using available training data, a forecast based on a new \( I \)-dimensional vector of predictor PCs \( x_0 \) is achieved through the \( J \) simultaneous regressions

\[
\hat{w} = R^T v_0 = R^T A^T x_0,
\]

where the \( J \) columns of the square regression matrix \( R \) contain the coefficients for the \( J \) regressions, the estimation of which is also method dependent. Since both predictors and predictands have been centered, there are no “intercept” terms in the regressions of Eq. (6). The vector mean forecast for the dimensional predictand anomalies can then be resynthesized as

\[
\hat{T}' = E_y (B^T)^{-1} \hat{w} = E_y (B^T)^{-1} R^T A^T x_0.
\]

This mean vector, combined with the forecast error covariance matrix \( S \), defined in the appendix for each of the three methods, specifies a full multivariate Gaussian probability distribution for the forecast anomalies. In the following, only univariate Gaussian forecast distributions for individual gridbox predictand temperatures will be considered, however, so that only the diagonal elements of \( S \) will be used.

1) CANONICAL CORRELATION ANALYSIS

Introduced to the atmospheric-sciences literature by Glahn (1968), CCA is by now a well-known method for finding maximally correlated projections onto coupled patterns in pairs of multivariate datasets (e.g., Bretherton et al. 1992; Wilks 2011). The computations may be initiated through the singular value decomposition of a scaled version of the cross-covariance submatrix \( S_{xy} \):

\[
S_{xx}^{-1/2} S_{xy} S_{yy}^{-1/2} = E_A R_{CCA} E_B^T.
\]

Here, \( S_{xx}^{-1/2} \) and \( S_{yy}^{-1/2} \) are symmetric matrices satisfying \( S_{xx}^{-1/2} S_{xx}^{-1/2} = S_{xx}^{-1} \) and \( S_{yy}^{-1/2} S_{yy}^{-1/2} = S_{yy}^{-1} \), the \( J \) columns of the matrices \( E_A \) and \( E_B \) contain vectors pertaining respectively to the predictors and predictands, and the \((J \times J)\) diagonal matrix \( R_{CCA} \) contains the canonical correlations. For CCA the pattern matrices in Eq. (5) are computed as

\[
A_{CCA} = S_{xx}^{-1/2} E_A \quad \text{and} \quad B_{CCA} = S_{yy}^{-1/2} E_B,
\]

and the regression matrix in Eqs. (6) and (7) is the diagonal matrix of canonical correlations, \( R = R_{CCA} \). To suppress the effects of sampling errors, regression
coefficients (i.e., the individual canonical correlations) in $R_{CCA}$ are set to zero in the following if they are not nominally significant at the 1% level, as was done also in Wilks (2008).

2) **MAXIMUM COVARIANCE ANALYSIS**

Rather than finding linear combinations in each of two datasets that maximize correlation, MCA instead seeks to maximize the covariances subject to an orthonormality constraint on the pattern vectors (e.g., Bretherton et al. 1992; Wilks 2011). These pairs of linear combinations are found through singular value decomposition of the unscaled cross-covariance submatrix $S_{xy}$:

$$ S_{xy} = A_{MCA} \Omega B_{MCA}^T. \quad (10) $$

Equation (10) is analogous to Eq. (8) for CCA, but here the left singular vectors that are the columns of the matrix $A_{MCA}$ are used to transform the predictor PCs [Eq. (5a)] and the right singular vectors in the matrix $B_{MCA}$ are used to transform the predictand PCs [Eq. (5b)]. The $(J \times J)$ diagonal matrix $\Omega$ containing the maximized covariances for the $J$ pattern pairs is not used in the forecasting procedure.

In CCA, the pairs of linear combinations in the elements of the vectors $v$ and $w$ [Eq. (5)] are uncorrelated with the elements of other pairs, and therefore the regression matrix $R_{CCA}$ used for prediction [Eq. (6)] is diagonal, implying that each of the $J$ regression may be considered independently of the others. This independence property does not hold for MCA, so that each column of the $(J \times J)$ MCA regression matrix $R_{MCA}$, for predicting the corresponding $j$th element of $w$, may include nonzero values for any of the predictor elements in $v$. Here, the coefficients in each column have been estimated in each case using a backward elimination predictor selection strategy, beginning with all predictors in the regression equation and successively removing the weakest until all are significant at the nominal 1% level. For the MCA forecasts produced below, the result was often but not always a diagonal matrix of MCA regression coefficients $R_{MCA}$.

The MCA forecasts are computed using Eq. (6), with $R = R_{MCA}$ and $A = A_{MCA}$. The forecast mean dimensional predictand anomalies are then resynthesized using Eq. (7), with $R = R_{MCA}$ and $A = A_{MCA}$, as well as $B = B_{MCA}$ from Eq. (10).

3) **REDUNDANCY ANALYSIS**

Both CCA and MCA are symmetric in their treatment of the predictors and predictands, in the sense that the derived pattern matrices and linear combinations are the same regardless of which of the two fields is considered to be which. An alternative derivation for the projection patterns is provided by RA, which specifically seeks those maximizing the predictand linear combination variances conditional on the predictors (e.g., Tippett et al. 2008; Tyler 1982; von Storch and Zwiers 1999).

For RA the columns of the predictand pattern matrix $B_{RA}$ are obtained as the eigenvectors of the real, symmetric $(J \times J)$ conditional covariance matrix for the predictand vector given the predictors

$$ S_y = S_{yx} S_{xx}^{-1} S_{xy}, \quad (11) $$

and each of the columns of $A_{RA}$ are then most conveniently computed using

$$ a_j = \lambda_j^{-1/2} S_{xx}^{-1} S_{xy} b_j \quad (12) $$

for the $j$th column, where $\lambda_j$ is the $j$th eigenvalue of the matrix in Eq. (11), and $b_j$ is its $j$th eigenvector (i.e., the $j$th column of $B_{RA}$).

In common with CCA, the $J$ regressions in Eq. (6) can be considered independently of each other, so that the RA regression matrix $R_{RA}$ is diagonal, with elements that are the square roots of the corresponding eigenvalues of the matrix in Eq. (11),

$$ R_{RA} = \Lambda^{1/2} = (B_{RA}^T S_{yx} A_{RA})^{1/2}. \quad (13) $$

Again, regression coefficients that are not nominally significant at the 1% level are set to zero in the following.

3. **Experimental setup**

a. **Data**

Predictor data have been derived from the extended reconstructed SST (ERSST.v3b) gridded dataset (Smith et al. 2008; available online at www.ncdc.noaa.gov/ersst/#grid). Seasonally averaged (i.e., running 3-month mean) SSTs for January 1894–February 2013 over the Indo-Pacific basin (13°S–55°N, 51°E–81°W) have been used. The data are serially complete, including reconstructions for missing values. These data are known to be less accurate in the earlier part of the record, but previous work (Wilks 2008) has found that use of the longer training data series nevertheless leads to substantially better seasonal predictions. The original 2° × 2° data have been averaged onto a 4° × 4° grid to reduce the computational overhead while retaining the large-scale SST features that are expected to provide the seasonal predictability, yielding 708 SST grid boxes. The leading PCs for these data are extracted, as described in section 3b, for use as predictors in the statistical forecasting procedures.
The predictand data were derived from the $5^\circ \times 5^\circ$ gridded North American ($25^\circ$-$65^\circ$N, $65^\circ$-$155^\circ$W) land surface temperatures in the Global Historical Climatology Network dataset (Peterson and Vose 1997). Version 3 of this dataset is available online (www.ncdc.noaa.gov/temp-and-precip/ghen-gridded-products.php) and includes corrections for data inhomogeneities (Menne and Williams 2009). Seasonal (3-month running mean) values for January 1895–May 2013 have been used, with any seasonal average that includes a missing month considered to be missing. Grid boxes having at least 90% of months not missing for this period were retained, yielding 58 locations primarily over the conterminous United States, southern Canada, and along the northern Pacific coast (Fig. 1). The predictand data are reduced to United States, southern Canada, and along the northern yielding 58 locations primarily over the conterminous of months not missing for this period were retained, considering to be missing. Grid boxes having at least 90% values for January 1895–May 2013 have been used, with any seasonal average that includes a missing month considered to be missing. Grid boxes having at least 90% of months not missing for this period were retained, yielding 58 locations primarily over the conterminous United States, southern Canada, and along the northern Pacific coast (Fig. 1). The predictand data are reduced to leading PCs for the statistical prediction procedures, as described in section 3b, and dimensional gridpoint forecast values have been resynthesized before calculation of verification metrics.

b. Experimental protocol

Temperature hindcasts for the grid boxes indicated in Fig. 1 have been computed for periods from January–March (JFM) 1981 through March–May 2013, using a protocol that emulates procedures that could have been carried out operationally during that period. In particular, the predictand value for a given year, or for any subsequent year, is not used in any way to formulate the forecast pertaining to it. Forecasts for lead times of 0 months [e.g., SST predictor PCs for October–December (OND) 1980 used to forecast JFM 1981] through 9 months (e.g., JFM 1980 to forecast JFM 1981) have been computed using each of the three methods described in section 2b. Predictor PCs have been recomputed for each forecast year to reflect data that would have become available most recently so that, for example, the training period OND 1894–1980 was used to forecast JFM 1981 at zero lead ($n = 87$ training years), OND 1894–1981 was used to forecast JFM 1982 at zero lead ($n = 88$ training years), and so on. These principal component analyses (PCAs) were based on covariance matrices centered using the entire available $n$-yr training period, with the SST anomalies weighted using the square root of the cosine of the latitudes (so that the variances being analyzed are scaled according to the cosines of the latitudes; North et al. 1982; Wilks 2011) to avoid unduly emphasizing the smaller higher-latitude grid boxes.

Predictand PCs were similarly recomputed each year to reflect the additional available year of training data, with the PCAs again based on the covariance matrix, and with the predictand data centered using the lagging 30-yr averages $T_{g,	au}^{(30)}$ recomputed for each prediction year $t$ as indicated by Eq. (4) (with predictand anomalies for the training years 1896–1925 centered using the 1895–1924 averages). Missing predictand gridbox values in the training data were set to zero anomaly for computation of the PCs. For each forecast year, season, lead time, and forecast method, the numbers of predictor and predictand PCs have been chosen according to a leave-one-out cross validation using the entire $n$-yr training period in each case. That combination of numbers of predictor and predictand PCs, $I$ and $J$, respectively, minimizing the continuous ranked probability score (Matheson and Winkler 1976; Wilks 2011), assuming Gaussian predictive distributions and averaged over training years and predictand grid boxes, was chosen separately for each forecast under the constraint $3 \leq J \leq I \leq 10$.

Probabilistic forecast accuracy has been assessed using the ranked probability score (RPS; Epstein 1969; Wilks 2011) on the basis of the climatological terciles, because this three-class format is standard for operational seasonal forecast dissemination (e.g., Barnston et al. 2010; Peng et al. 2013; van den Dool 2007), and its use allows direct comparison of results with other published seasonal forecasting verifications. These climatological terciles have been updated annually for each season and grid box, again based on lagging 30-yr climatological statistics. In particular, the lower and upper terciles were recomputed for each target year $t$ as

$$L_{g,t} = s_{g,t}^{(30)} \Phi^{-1}(0.3333) + T_{g,t}^{(30)} = -0.4307 s_{g,t}^{(30)} + T_{g,t}^{(30)},$$

(14a)

and

$$U_{g,t} = s_{g,t}^{(30)} \Phi^{-1}(0.6667) + T_{g,t}^{(30)} = +0.4307 s_{g,t}^{(30)} + T_{g,t}^{(30)},$$

(14b)

respectively, where $\Phi^{-1}()$ denotes the quantile (i.e., inverse cumulative distribution) function of the standard Gaussian distribution, $T_{g,t}^{(30)}$ is the $g$th predictand gridbox.
mean for target year \( t \) [i.e., the \( g \)th element of the vector \( \mathbf{T}^{(30)}_t \) in Eq. (4)], and the lagging 30-yr climatological standard deviations are

\[
\hat{s}^{(30)}_{g,t} = \left\{ \frac{1}{29} \sum_{\tau=1-30} \left[ T_{g,\tau} - \mathbf{T}^{(30)}_g \right]^2 \right\}^{1/2}. 
\]

(15)

c. Trend adjustment methods

The methods described in section 2 produce forecasts of conditional mean anomalies \( \mathbf{T} \) [Eq. (7)] and corresponding forecast error variances \( \mathbf{S}_T \) (as described in the appendix). Together these are sufficient to define forecast Gaussian probability distributions for the gridbox anomalies. Three methods for constructing the probability forecasts for the actual predictand temperatures from these distributions are described in this section.

The default approach to constructing predictive distributions for the gridbox predictand temperatures is to shift the gridbox forecast anomaly distributions by adding the lagging 30-yr climatological average,

\[
\mathbf{T}^{(30)}_{g,t} = \mathbf{T}'_{g,t} + \mathbf{T}^{(30)}_g.
\]

(16)

Together with the corresponding forecast standard deviations \( \hat{s}_{g,t} \), which are the square roots of the diagonal elements of \( \mathbf{S}_T \), the three-category probability forecasts are constructed as

\[
\hat{p}_{1,g}^{(30)} = \Phi\left\{ \left[ L_{g,t} - \mathbf{T}^{(30)}_{g,t} \right]/\hat{s}_{g,t} \right\},
\]

\[
\hat{p}_{2,g}^{(30)} = \Phi\left\{ \left[ U_{g,t} - \mathbf{T}^{(30)}_{g,t} \right]/\hat{s}_{g,t} \right\} - \Phi\left\{ \left[ L_{g,t} - \mathbf{T}^{(30)}_{g,t} \right]/\hat{s}_{g,t} \right\},
\]

and

\[
\hat{p}_{3,g}^{(30)} = 1 - \Phi\left\{ \left[ U_{g,t} - \mathbf{T}^{(30)}_{g,t} \right]/\hat{s}_{g,t} \right\}
\]

(17)

for each target year \( t \), where the subscripts 1, 2, and 3 indicate the cool, near-normal, and warm predictand categories, respectively. In cases in which Eq. (17) yielded probabilities \( \hat{p}_{2,g}^{(30)} \) for the near-normal category that were larger than the climatological value of \( 1/3 \), the forecast standard deviation \( \hat{s}_{g,t} \) has been increased until \( \hat{p}_{2,g}^{(30)} = 1/3 \). The effect of this adjustment is to allow predictive distributions exhibiting high confidence (low standard deviation) only in cases in which the mean forecast also exhibits high confidence (i.e., differs substantially from its climatological value). Tercile forecasts favoring the middle category are expected a priori to perform poorly (van den Dool and Toth 1991), as has been seen for both seasonal (e.g., Barnston et al. 2010; Wilks 2000; Wilks and Godfrey 2002) and medium-range (Wilks 2013b) forecasts.

Even though the 30-yr lagging climatological averages are updated annually, this baseline for the forecast anomalies has failed to keep up with the ongoing climate warming during the past several decades (Livezey et al. 2007; Wilks 2013a; Wilks and Livezey 2013), and therefore forecasts that are based on Eq. (16) are expected to exhibit strong negative (cold) biases. Better but still cold-biased projections of gradually shifting seasonal climate means can be obtained using the lagging 15-yr averages

\[
\mathbf{T}^{(15)}_{g,t} = \frac{1}{15} \sum_{\tau=1-15} T_{g,\tau},
\]

(18)

which have performed best overall to date for seasonal U.S. temperature averages in terms of minimizing mean square error and bias, among a variety of alternatives investigated (Wilks 2013a; Wilks and Livezey 2013). Analogous to Eqs. (16) and (17), Gaussian forecast means and corresponding three-category probability forecasts can be constructed on the basis of this more nimble projection of the climate mean:

\[
\hat{T}^{(15)}_{g,t} = \hat{T}'_{g,t} + \mathbf{T}^{(15)}_g \quad \text{and} \quad \hat{p}_{1,g}^{(15)} = \Phi\left\{ \left[ L_{g,t} - \hat{T}^{(15)}_{g,t} \right]/\hat{s}_{g,t} \right\},
\]

\[
\hat{p}_{2,g}^{(15)} = \Phi\left\{ \left[ U_{g,t} - \hat{T}^{(15)}_{g,t} \right]/\hat{s}_{g,t} \right\} - \Phi\left\{ \left[ L_{g,t} - \hat{T}^{(15)}_{g,t} \right]/\hat{s}_{g,t} \right\}, \quad \text{and}
\]

\[
\hat{p}_{3,g}^{(15)} = 1 - \Phi\left\{ \left[ U_{g,t} - \hat{T}^{(15)}_{g,t} \right]/\hat{s}_{g,t} \right\}.
\]

(19)

As before, the condition \( \hat{p}_{2,g}^{(15)} \leq 1/3 \) has been enforced by inflating the forecast standard deviation \( \hat{s}_{g,t} \) if necessary.

An alternative for updating the tercile forecast probabilities to reflect ongoing climate trends has recently been proposed by Krakauer et al. (2013). Their naive Bayesian posterior (NBP) probabilities are obtained by updating trend-following climatological probabilities \( \hat{p}_{k,g}^{(15)}, k = 1, 2, 3 \), using the default forecast probabilities from Eq. (17), through Bayes’s theorem,

\[
\hat{p}_{k,g}^{(NBP)} = \frac{\hat{p}_{k,g}^{(30)} \hat{p}_{k,g}^{(15)}}{\sum_{l=1}^{3} \hat{p}_{l,g}^{(30)} \hat{p}_{l,g}^{(15)}},
\]

(20)

where the trend-following climatological probabilities are taken here to be based on the lagging 15-yr means,

\[
\hat{p}_{k,g}^{(15)} = \Phi\left\{ \left[ L_{g,t} - \mathbf{T}^{(15)}_{g,t} \right]/\hat{s}_{g,t}^{(30)} \right\},
\]

and...
tributions, following Briggs and Wilks (1996), yielding them to be composed of three truncated Gaussian dis-
temperatures. Means for the NBP probability-adjusted the imputed NBP forecast means

For the lagging 30- and 15-yr averages, the biases are higher, especially for the MCA forecasts. In a similar way, these biases are typically also smaller than those for the climatological lagging 15-yr means (rightmost bars at the right in Fig. 2). The RPS skills for the lagging 15-yr and NBP forecasts are generally similar, although the biases for the forecasts centered on the lagging 15-yr means are smaller, with 90% confidence intervals that often cover zero. These biases are also smaller than those for the climatological lagging 15-yr means (rightmost bars at the right in Fig. 2), whereas the biases for the NBP forecasts are similar to these 15-yr climatological biases.

Table 1 shows RPS skills for the forecasts aggregated over all seasons, using the same lead-time stratifications as in Fig. 2. Again the lagging 15-yr and NBP trend-adjustment methods perform similarly, with slightly higher skills for the NBP probability adjustments. Results using the annually updated 30-yr averages are clearly weaker. Previously reported RPS skills for the U.S. Climate Prediction Center (CPC) seasonal temperature forecasts at the shortest lead times for 1995–2002 (Higgins et al. 2004), 1995–2009 (Peng et al. 2012, 2013), and 1995–98 (Wilks 2000) were approximately 0.03. These older skill scores were computed with respect to the baseline climatological values [Eq. (4)], and the lagging-15-yr and NBP results, which do. The RPS skill is lowest for the lagging 30-yr forecast means in each of the 36 comparisons with the two trend adjustment approaches. Concurrently, the biases exhibited by the lagging-30-yr means are consistently larger (more negative, or colder) than for the other approaches, although they generally improve upon the biases of the climatological lagging 30-yr means (left gray bars at the right in Fig. 2).

The RPS skills for the lagging 15-yr and NBP forecasts are generally similar, although the biases for the forecasts centered on the lagging 15-yr means are smaller, with 90% confidence intervals that often cover zero. These biases are typically also smaller than those for the climatological lagging 15-yr means (rightmost bars at the right in Fig. 2), whereas the biases for the NBP forecasts are similar to these 15-yr climatological biases.

4. Results

Figure 2 shows the main comparative results for forecast skill and bias. Each of the four panels in Fig. 2 aggregates results for three consecutive target seasons to conserve space and smooth some of the sampling variations. As indicated at the bottom of Fig. 2, lead times are stratified into the three groups 0–1, 2–5, and 6–9 months; within each of which the groups of three bars pertain to forecast distributions centered using the lagging 30-yr averages [Eq. (16)], lagging 15-yr averages [Eq. (19)], and the NBP probability adjustments [Eq. (21)], as indicated at the left of Fig. 2a. Each individual group of three bars shows results for CCA (left), MCA (middle), and RA (right) forecasts. The black-outlined bars show RPS skill for the forecasts computed in the conventional way (e.g., Wilks 2011), using the climatological reference forecast of equal probabilities for the three categories, which were defined in Eq. (14) on the basis of the annually updated 30-yr climatological distribution. The average forecast biases are shown as the gray-outlined bars. In all cases the whiskers show 90% bootstrap confidence intervals, for which the computations have respected the spatial coherency of the predictand fields by resampling entire simultaneous spatial fields rather than individual grid-box values.

For the lagging 30- and 15-yr averages, the biases are simply the average differences between the mean forecasts [Eqs. (16) and (19)] and the corresponding observed temperatures. Means for the NBP probability-adjusted forecasts [Eq. (21)] have been calculated by considering them to be composed of three truncated Gaussian distributions, following Briggs and Wilks (1996), yielding the imputed NBP forecast means

\[
\mathcal{P}^{(15)}_{g, t} = \Phi\{[U_{g, t} - \mathcal{T}^{(15)}_{g, t}]\delta^{(30)}_{g, t}\} \\
- \Phi\{[L_{g, t} - \mathcal{T}^{(15)}_{g, t}]\delta^{(30)}_{g, t}\},
\]

and

\[
\mathcal{P}^{(15)}_3 = 1 - \Phi\{[U_{g, t} - \mathcal{T}^{(15)}_{g, t}]\delta^{(30)}_{g, t}\}. \tag{22}
\]

which follows because the conditional means of the upper and lower thirds of the standard Gaussian distribution are \(\pm 1.0908\).

Forecast skill is generally highest for the spring (Fig. 2b) and summer (Fig. 2c) seasonal aggregations. The fall seasonal aggregation (Fig. 2d) exhibits the weakest, although still generally positive, skill. Except for the fall seasonal aggregation, the skill generally declines with increasing lead time, but often only weakly so, which suggests that much of the skill may have been derived from capturing trends (Livezey and Timofeyeva 2008). The three forecasting methods generally perform similarly to each other, with respect to both RPS skill and bias, although there is a tendency for the MCA forecasts to exhibit slightly higher skill than CCA or RA. In a majority (28 of 36) of comparisons the MCA forecasts exhibit the highest skill of the three, which cannot be explained by a corresponding reverse pattern in the respective biases.
Fig. 2. RPS skill (black bars) and forecast bias (gray bars) for the seasonal groupings (a) winter [November, December, and January (NDJ–JFM)], (b) spring (FMA–AMJ), (c) summer [May, June, and July (MJJ–JAS)], and (d) autumn [August, September, and October (ASO–OND)]. Left, middle, and right groups of nine bars stratify the lead times 0–1, 2–5, and 6–9 months, respectively; within each of these groups of nine the left, middle, and right groups of three bars show results for lagging 30-yr forecast means [Eq. (16)], lagging 15-yr forecast means [Eq. (19)], and the NBP probability adjustment [Eq. (21)]. Individual bars within each group of three show results for CCA, MCA, and RA forecasts. Gray bars to the extreme right of each panel show biases for the lagging 30- and 15-yr climatological values only. Whiskers in each case show the 90% bootstrap confidence intervals.
temperature forecasts over North America at zero lead time, for 1997–2000. Little seasonal or lead-time dependence was observed for the numbers of predictor and predictand PCs, I and J, chosen by the cross-validation training procedure, although there was substantial year-to-year variability in these optimized parameters. In the overwhelming majority of cases, seven or fewer predictor PCs were chosen, with as many or nearly as many predictand PCs, so that nearly always $J \geq I - 2$. For CCA forecasts the most commonly chosen combinations were $I = J = 3$ and $I = J = 4$. For MCA the most common combinations were $I = J = 3$, $I = J = 4$, $I = J = 5$, and $I = J = 6$. For RA the most common combination was $I = J = 3$.

Wilks (2008) found that substantially better CCA and MCA forecasts were achieved using long (nineteenth century onward) training series rather than more conventional shorter, modern (1951 onward) predictand training data only. Figure 3 shows results for the spring [February, March, and April (FMA)--April, May, and June (AMJ)] seasonal aggregation, comparable to Fig. 2b, when only predictand training data from 1951 onward have been used, again to forecast for the years 1981–2013. As in the previous work the RPS skills are notably weaker than those for the longer training period shown in Fig. 2b, for all three forecasting methods and all three mean adjustment approaches, although for these shorter training data the MCA forecasts yield consistently more skillful forecasts in all nine comparisons. The spring-season results in Fig. 3 are representative of relative forecast performance using the shorter training series for the other seasonal stratifications as well.

Figure 4 illustrates the calibration of the Gaussian predictive distributions using probability integral transform (PIT) histograms (Gneiting et al. 2005; Wilks 2011) for the 22 564 nonmissing FMA–AMJ forecasts at 2–5-month lead times, which again are representative. The PIT histogram is the extension for continuous predictive distributions of the more familiar verification rank histogram for ensemble forecasts and can be interpreted using the same diagnostics (Hamill 2001; Wilks 2011). Observations are classified into one of the 20 bins on the horizontal axis according to the cumulative probability locations in their corresponding predictive distributions,

$$
\Phi\left(\frac{T_{g,t} - \bar{T}_{g,t}}{\hat{\sigma}_{g,t}}\right),
$$

with the histogram bars tabulating numbers of observations in each bin. For perfectly calibrated Gaussian predictive distributions, these histograms will conform to the horizontal dashed lines, within allowance for sampling variations. Outlined open bars in Fig. 4 show PIT histograms for forecasts in which the constraint $\hat{p}_{2g,t} \leq \frac{1}{3}$ has been enforced, gray bars show results for the

### Table 1: Average RPS skill relative to annually updated 30-yr climate normals, aggregating forecasts for all seasons, for the indicated lead times (months).

<table>
<thead>
<tr>
<th>Lead</th>
<th>CCA</th>
<th>MCA</th>
<th>RA</th>
<th>CCA</th>
<th>MCA</th>
<th>RA</th>
<th>CCA</th>
<th>MCA</th>
<th>RA</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–1</td>
<td>0.0203</td>
<td>0.0254</td>
<td>0.0177</td>
<td>0.0406</td>
<td>0.0525</td>
<td>0.0437</td>
<td>0.0431</td>
<td>0.0534</td>
<td>0.0464</td>
</tr>
<tr>
<td>2–5</td>
<td>0.0214</td>
<td>0.0213</td>
<td>0.0169</td>
<td>0.0429</td>
<td>0.0511</td>
<td>0.0440</td>
<td>0.0453</td>
<td>0.0517</td>
<td>0.0457</td>
</tr>
<tr>
<td>6–9</td>
<td>0.0077</td>
<td>0.0153</td>
<td>0.0068</td>
<td>0.0323</td>
<td>0.0456</td>
<td>0.0350</td>
<td>0.0342</td>
<td>0.0463</td>
<td>0.0371</td>
</tr>
</tbody>
</table>

**Fig. 3.** As in Fig. 2b, but for FMA–AMJ forecasts trained using data from 1951 onward, only.
unadjusted Gaussian predictive distributions, and the legends indicate the $\chi^2$ lack-of-fit (relative to the uniform distributions indicated by the horizontal dashed lines) statistics in each case.

All panels in Fig. 4 exhibit upward slopes of the PIT histogram bars, which is diagnostic for underforecasting (cold) biases, as expected. In comparing individual case pairs between Figs. 4a and 4b, however, it is evident that use of the lagging 15-yr means substantially reduces these biases, consistent with the bias results shown in Fig. 2. The gray bars, showing results for which the middle-category constraint $\hat{p}_{2,g} \leq 1/3$ has not been enforced, exhibit in addition the tendency for U-shaped PIT histograms, which is diagnostic for underdispersion, or overconfidence. Thus, constraining $\hat{p}_{2,g} \leq 1/3$ improves calibration, especially for the CCA and RA forecasts. This improvement is also evident in the resulting decreases in the $\chi^2$ statistics for the outlined open bars in each of the six comparisons. In a similar way, use of lagging 15-yr means rather than 30-yr means decreases $\chi^2$ in each comparison, and the MCA forecasts yield better calibrated predictive distributions (smaller $\chi^2$) than do either CCA or RA. Best calibrated forecasts overall are produced by MCA, together with the lagging 15-yr means and the $\hat{p}_{2,g} \leq 1/3$ constraint. Not surprising given the cold biases present in all the forecasts is that the PIT histograms in Fig. 4 all exhibit significant deviations from uniformity, as the 99.99th
percentile of the $\chi^2$ distribution with 19 degrees of freedom is 50.8.

Figure 5 shows calibration simplexes (Wilks 2013b). The calibration simplex is an extension of the more familiar reliability diagram for probability forecasts of yes–no predictands (Murphy and Winkler 1977; Wilks 2011). Again results are shown for FMA–AMJ forecasts at 2–5-month lead times, formulated using 30-yr lagging means (Fig. 5a), 15-yr lagging means (Fig. 5b), and the NBP probabilities (Fig. 5c). Each hexagon in the calibration simplex represents one of the possible three-element probability vectors, when the probabilities have been rounded to integer multiples of $\frac{1}{9}$. Probabilities for the cool outcome increase to the left along the bottoms of the simplexes, probabilities for the near-normal outcome increase upward along the left edges, and probabilities for the warm outcome increase downward along the right edge. The sizes of the plotting symbols indicate the subsample sizes for each of the combinations of forecast probabilities, and only subsample sizes larger than 20 have been plotted. The absence of plotting symbols in the top halves of the simplexes in Figs. 5a and 5b reflects the $p_{2.0} \leq {1/3}$ constraint. The placement of the symbol within each hexagon shows forecast calibration.
for the respective probability vector, with correctly calibrated forecasts indicated by symbols at the centers of the hexagons. There is a clear tendency for displacement of the plotting symbols toward the “A” corner, which is diagnostic for insufficient probability assigned to the warm outcome (i.e., cold bias), again consistent with results in Figs. 2 and 4.

The most prominent differences in Fig. 5 are between the forecasts formulated using the lagging 30-yr means (Fig. 5a) and those that adjust for trend using the lagging 15-yr means (Figs. 5b,c). In the former case the climatological forecast consisting of equal probabilities for each of the three categories is overwhelmingly the most frequently used and simultaneously exhibits strong cold bias. In Figs. 5b and 5c the climatological forecast $\frac{1}{3}-\frac{1}{3}-\frac{1}{3}$ is used less frequently than the warmer forecast $\frac{2}{3}-\frac{1}{3}-\frac{1}{3}$, with both of these forecast subsamples exhibiting better calibration for all three forecast methods.

5. Discussion and conclusions

This paper has compared three linear multivariate statistical forecasting tools—CCA, MCA, and RA—for constructing Probabilistic seasonal forecasts of grid box North American temperatures. The RPS skill of these forecasts compares favorably to previously published results for North American seasonal forecasts (e.g., Barnston et al. 2010; Higgins et al. 2004; Peng et al. 2012, 2013; Wilks 2000; Wilks and Godfrey 2002), even though the annually updated lagging 30-yr averages used here have imposed a more challenging skill baseline for the present forecasts to improve upon. Forecast accuracy has been reported in terms of RPS skill so that these comparisons could readily be made, but characterizing probabilistic forecast skill in terms of continuous RPS, or skill of the mean forecast in terms of mean-square-error skill (not shown), yields the same conclusions as in Figs. 2 and 3 about relative performance of the methods for the different seasons, lead times, and training-data lengths.

Overall, the three methods exhibited generally similar skill levels, consistent with the results of Tippett et al. (2008), who compared their use in statistical downscaling. There was, however, a tendency for the MCA forecasts to be somewhat more skillful when trained on the full (1895 onward) predictand data record, as well as yielding better forecast calibration. MCA forecasts were decisively preferred when trained on a shorter (1951 onward) record. For all three forecasting methods, constraining the forecast probability for the “near normal” category to be no larger than its climatological value of $\frac{1}{3}$ improved forecast calibration and skill. Consistent with previous results (Wilks 2008), use of the shorter training period provided clearly less accurate forecasts, including the case of RA forecasts for which the sensitivity to training-period length has not previously been investigated.

Another purpose of this study has been to investigate different approaches to incorporating the effects of recent climate warming into the forecast procedure to reduce the tendency for the cold biases that this warming tends to generate, and thereby also to increase skill. Consistent with recent results (Wilks 2013a; Wilks and Livezey 2013), adding the lagging 15-yr mean to the forecast temperature anomalies [similar to Higgins et al. (2004), who used 11-yr means] substantially increased the forecast skill and reduced the cold bias relative to the more conventional lagging 30-yr means. There appeared to be a small skill advantage for the recently introduced (Krakauer et al. 2013) NBP probability adjustment over the simpler mean adjustment, although the NBP method yielded somewhat worse cold biases. Although the lagging 15-yr mean to date has been the best overall method for short-term North American seasonal temperature mean projections, other more sophisticated alternatives (Wilks 2013a; Wilks and Livezey 2013) may emerge as being preferable in coming years as the ongoing climate warming unfolds, or for longer-lead (e.g., decadal) predictions.

The appendix provides for the first time expressions for the MCA and RA forecast error variances and covariances, similar to the CCA forecast error covariance expressions in Wilks (2014). Together with the conditional mean forecasts that are the conventional outputs of forecasts derived from the three methods, these forecast error variances define Gaussian forecast distributions for the predictand anomalies. For seasonal averages of a statistically well-behaved quantity such as temperature, the central limit theorem leads to the expectation that the Gaussian distribution will provide an excellent statistical model for the predictive distributions. For other predictands, such as seasonal precipitation totals, transforming the predictands to a near-Gaussian distribution may be advisable before completing the computations. Although only univariate predictive distributions for individual grid boxes were evaluated here, these methods could also be used to formulate joint multivariate normal predictive distributions for all grid boxes simultaneously, for chosen subsets of grid boxes, or for linear combinations of gridbox temperatures (e.g., regional averages).

In addition to providing the basis for probabilistic forecasting with CCA, MCA, and RA, the forecast error covariance matrices extend the range of uses for these methods to other applications. Climate mean fields that have been statistically downscaled using one of these
methods (e.g., Tippett et al. 2008) could be used in conjunction with their covariance matrices to stochastically simulate sequences of individual climate fields varying coherently around the downscaled means. In a similar way, stochastic ensembles of statistical ocean realizations forcing dynamical atmospheres (e.g., Goddard and Graham 1999) or ensembles of statistical atmospheres forcing dynamical oceans (e.g., Barnett et al. 1993) could also be generated.

Acknowledgments. This research was supported by the National Science Foundation under Grant AGS-1112200.

APPENDIX

Prediction Covariance Matrices

Equation (7) specifies a forecast mean vector for temperature anomalies at the gridbox locations indicated in Fig. 1. Since the predictands are long (i.e., seasonal) time averages, their distribution is expected, under the central limit theorem, to be multivariate Gaussian or nearly so. Therefore, calculation of a forecast error covariance matrix completes the specification of the Gaussian predictive probability distributions.

For the CCA forecasts, Wilks (2014) has derived this error covariance matrix. Because each of the J regressions in CCA can be considered independently, the forecast error covariance for the predictand projection vector \( \mathbf{w} \) [Eq. (6)] is diagonal:

\[
\mathbf{S}_w = \text{diag} \left( 1 - r_{C,j}^2 \left( 1 + \frac{1}{n} \sum_{m=1}^{n} \frac{v_{0j}^2}{v_{mj}^2} \right) \right),
\]

where \( r_{C,j} \) is the jth canonical correlation, \( 1 - r_{C,j}^2 \) is the estimated error variance for the jth regression, \( v_{mj}^2 \) is the jth element of the predictor projection vector for training year \( m \), and \( v_{0j}^2 \) is the corresponding predictor element on which the current forecast is conditioned. Equation (A1) is the consequence of a well-known result for univariate regression (e.g., Draper and Smith 1981; Wilks 2011). Since the dimensional forecast temperature anomaly vector [Eq. (7)] is simply a linear transformation of Eq. (6), its covariance matrix is computed from Eq. (A1) as

\[
\mathbf{S}_\mathbf{f} = \mathbf{S}_w \mathbf{B}_{\mathbf{CCA}}^{-1} \mathbf{E}_\mathbf{f}^\mathsf{T}.
\]

Computing the predictand error covariance matrix in MCA is somewhat more complicated, because any of the elements of the predictor projection vector \( \mathbf{v} \) may be used in a regression for any of the elements of the predictand projection vector \( \mathbf{w} \). That is, Eq. (6) is a multivariate regression (Johnson and Wichern 2002), in which the regression matrix \( \mathbf{R}_{\mathbf{MCA}} \) may not be diagonal:

\[
\hat{\mathbf{w}} = \mathbf{R}_{\mathbf{MCA}}^\mathsf{T} \mathbf{v} + \mathbf{e}.
\]

Here, each of the columns of the \((J \times n)\) matrices \( \hat{\mathbf{w}}, \mathbf{v}, \) and \( \mathbf{e} \) contains one of the vectors \( \hat{\mathbf{w}}, \mathbf{v}, \) and \( \mathbf{e} \) for the \( n \) training years. The \( \mathbf{e} \) vectors contain estimated regression residuals for a given year. The forecast error covariance for the predictand projection vector can be computed as (Johnson and Wichern 2002)

\[
\mathbf{S}_w = \frac{1 + \mathbf{v}_0^\mathsf{T} (\mathbf{v} \mathbf{v}^\mathsf{T})^{-1} \mathbf{v}_0}{n - J - 1} \mathbf{e} \mathbf{e}^\mathsf{T},
\]

after which \( \mathbf{S}_\mathbf{f} \) can be calculated by substituting into Eq. (A2) using \( \mathbf{B}_{\mathbf{MCA}} \) in place of \( \mathbf{B}_{\mathbf{CCA}} \).

For RA, each of the \( J \) regressions in Eq. (6) can be considered independently, as in CCA, again yielding a diagonal forecast error covariance matrix for the predictand projection vector,

\[
\mathbf{S}_w = \text{diag} \left[ -\lambda_j + \sum_{k=1}^{J} \gamma_k \lambda_k b_{kj}^2 \left( 1 + \frac{1}{n} \sum_{m=1}^{n} \frac{v_{0jm}^2}{v_{mjm}^2} \right) \right].
\]

Equation (A5) for RA differs from its counterpart Eq. (A1) for CCA in the calculation of the estimated residual variance (first parenthetical factor in the square brackets). Here, \( \gamma_k \) denotes the \( k \)th of the \( J \) leading eigenvalues associated with the eigenvectors in the columns of the matrix \( \mathbf{E}_\mathbf{f} \) [Eq. (3)], and \( b_{kj} \) denotes the \( k \)th element in the \( j \)th column of the matrix \( \mathbf{B}_{\mathbf{RA}} \) [eigenvectors of Eq. (11)]. Again, having computed \( \mathbf{S}_w, \mathbf{S}_\mathbf{f} \) can be calculated by substituting into Eq. (A2) using \( \mathbf{B}_{\mathbf{RA}} \) in place of \( \mathbf{B}_{\mathbf{CCA}} \).

REFERENCES


