Best Practices for Postprocessing Ensemble Climate Forecasts. Part I: Selecting Appropriate Recalibration Methods

PHILIP G. SANSOM, CHRISTOPHER A. T. FERRO, AND DAVID B. STEPHENSON

Exeter Climate Systems, University of Exeter, Exeter, United Kingdom

LISA GODDARD AND SIMON J. MASON

International Research Institute for Climate and Society, The Earth Institute, Columbia University, Palisades, New York

(Manuscript received 25 November 2015, in final form 9 May 2016)

ABSTRACT

This study describes a systematic approach to selecting optimal statistical recalibration methods and hindcast designs for producing reliable probability forecasts on seasonal-to-decadal time scales. A new recalibration method is introduced that includes adjustments for both unconditional and conditional biases in the mean and variance of the forecast distribution and linear time-dependent bias in the mean. The complexity of the recalibration can be systematically varied by restricting the parameters. Simple recalibration methods may outperform more complex ones given limited training data. A new cross-validation methodology is proposed that allows the comparison of multiple recalibration methods and varying training periods using limited data.

Part I considers the effect on forecast skill of varying the recalibration complexity and training period length. The interaction between these factors is analyzed for gridbox forecasts of annual mean near-surface temperature from the CanCM4 model. Recalibration methods that include conditional adjustment of the ensemble mean outperform simple bias correction by issuing climatological forecasts where the model has limited skill. Trend-adjusted forecasts outperform forecasts without trend adjustment at almost 75% of grid boxes. The optimal training period is around 30 yr for trend-adjusted forecasts and around 15 yr otherwise. The optimal training period is strongly related to the length of the optimal climatology. Longer training periods may increase overall performance but at the expense of very poor forecasts where skill is limited.

1. Introduction

Reliable projections of future climate on seasonal-to-decadal time scales have enormous potential value for adaptation purposes. In response to rising demand for such forecasts, the World Meteorological Organization designated 12 Global Producing Centres for Long Range Forecasts (WMO 2007). Decadal prediction experiments were also included in CMIP5, with results contributed by 16 institutions (Taylor et al. 2012). However, long-range forecasts are subject to large uncertainties arising from a variety of sources (e.g., initial conditions, model parameters, and model structure). Decision makers need to know the range of possible outcomes in order to make informed choices about adaptation measures, not just the most likely outcome. Therefore, probability forecasts that reliably quantify the uncertainty in long-range forecasts are required.

Dynamical forecast models are valuable tools for predicting future climate, but they are far from perfect. Climate models suffer from a variety of biases and errors that can lead to probabilistically unreliable forecasts. Therefore, forecasts on seasonal-to-decadal time scales benefit from statistical recalibration based on past performance. The term "bias correction" is often used interchangeably to refer to adjustments applied within the forecast model (e.g., flux corrections) or simply removing the mean difference between a set of historical forecasts and observations. This study only considers adjustments that are performed outside of the forecast model. The term "recalibration" is used to distinguish

Corresponding author address: Philip G. Sansom, Exeter Climate Systems, University of Exeter, Laver Building, North Park Road, Exeter EX4 4QE, United Kingdom.
E-mail: p.g.sansom@exeter.ac.uk

DOI: 10.1175/JCLI-D-15-0868.1

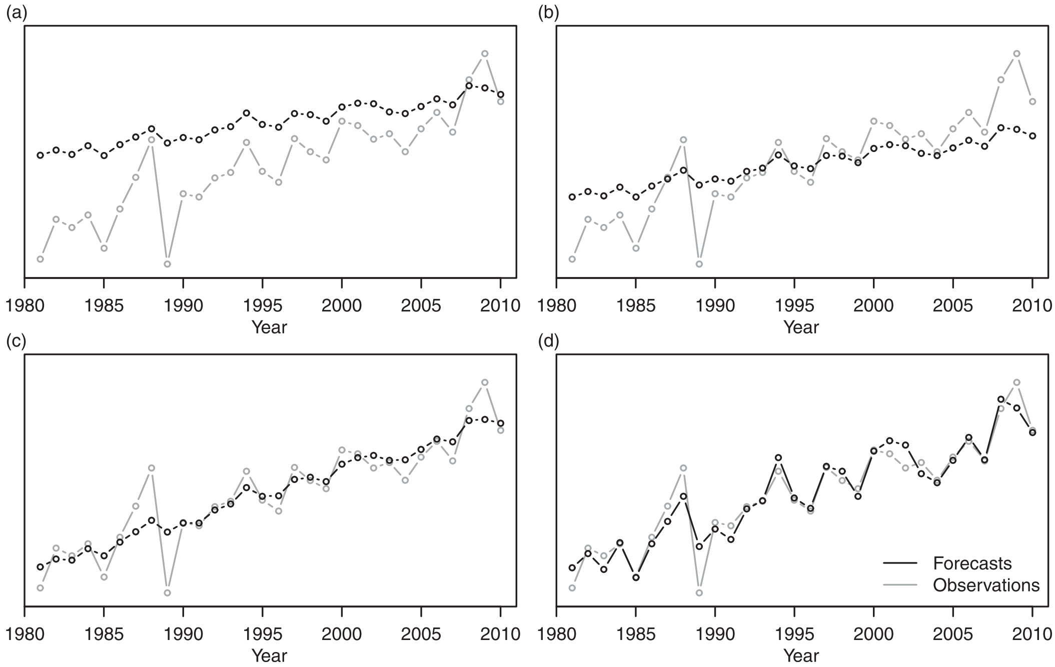

FIG. 1. Examples of the impact of different types of recalibration of the forecast ensemble mean. Synthetic time series of ensemble mean forecasts (black) and observations (gray) (a) before recalibration, (b) after unconditional adjustment, (c) after unconditional and trend adjustment, and (d) after unconditional, trend, and conditional adjustment.

these from both adjustments made inside the model, and from the practice of tuning model parameters (calibration). Recalibration can refer to any adjustment of the statistics of the forecast ensemble (e.g., the mean or variance).

A number of methods have been proposed in the literature for recalibrating probability forecasts. These include best member dressing (Roulston and Smith 2003), logistic regression (Hamill et al. 2004), Bayesian model averaging (Raftery et al. 2005), forecast assimilation (Stephenson et al. 2005), and ensemble model output statistics (Gneiting et al. 2005). Comparisons by Wilks (2006) and Williams et al. (2014) concluded that Bayesian model averaging, best member dressing, and ensemble model output statistics all perform similarly. Some of these methods also allow forecasts to be combined from multiple models simultaneously. Combining forecasts from multiple models adds an additional layer of complexity. Therefore, this study is restricted to methods for recalibrating forecasts from a single model. Excellent reviews of the issues surrounding multimodel combination are provided by Tebaldi and Knutti (2007), Knutti et al. (2010), and Stephenson et al. (2012).

Different studies have emphasized different types of recalibration for long-range forecasts. Unconditional biases are defined as differences between forecasts and observations that are constant between forecast times. Constant biases in the forecast means are routinely removed by the use of anomalies about a long-term mean (e.g., Stockdale 1997). Figure 1b shows the effect of removing a constant bias from the synthetic forecasts in Fig. 1a. Unconditional biases are usually interpreted as a discrepancy between the equilibrium state of a model and the Earth system.

Time-dependent biases can occur if the model response to anthropogenic climate change differs from that of the Earth system, or if there is a systematic error in the observations. Time-dependent biases in the forecast mean are often adjusted by removing a linear trend estimated from the forecasts and adding back a linear trend estimated from the observations (Kharin et al. 2012). Figure 1c shows the effect of linear trend adjustment on the bias-adjusted forecasts in Fig. 1b.

Conditional biases in the forecast mean are systematic errors in the strength of the predictable signal. Conditional biases in the forecast means can be minimized by linear scaling (Weigel et al. 2009; Eade et al. 2014). Figure 1d shows the effect of linear scaling on the trend-adjusted forecasts in Fig. 1c. Conditional biases in decadal forecasts can be so large that model forecasts are outperformed in mean-square error by climatological forecasts (i.e., the mean and variance of the observational record), even though the forecasts may be highly correlated with the observations (Goddard et al. 2013, their Figs. 5 and 6).

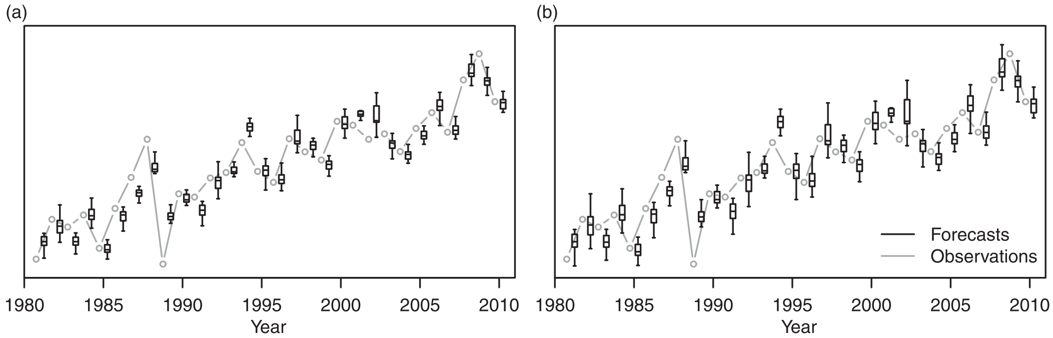

FIG. 2. Example of the impact of recalibration of the forecast variance. Synthetic time series of overconfident forecast ensembles after conditional, unconditional, and trend adjustment of the ensemble mean, (a) before and (b) after linear scaling of the ensemble variance. Box-and-whisker plots represent the spread of the forecast ensembles.

Different forms of recalibration are sometimes recommended depending on whether forecasts are initialized from full-field observations or anomalies (ICPO 2011). Full-field initialization can lead to sudden shocks or gradual drifts in individual forecasts that must be corrected based on historical performance. These effects are caused by differences between the equilibrium state of the model and the Earth system. Initialization by anomalies is intended to minimize these effects. In practice, anomaly initialized forecasts have also been found to benefit from recalibration based on past performance (Smith et al. 2013). Linear trend adjustment has also been applied to correct systematic drifts within individual forecasts in order to recalibrate multiple lead times simultaneously (Kharin et al. 2012). For a fixed lead time, systematic drifts within individual forecasts appear as unconditional biases. This study develops recalibration methods applicable to both full-field and anomaly initialized forecasts at fixed lead times.

To obtain reliable probability forecasts, the forecast uncertainty may also require adjustment. The forecast uncertainty is often estimated by the mean-square error of a set of recalibrated historical forecast means (e.g., Kharin and Zwiers 2003; Glahn et al. 2009). The mean-square error estimate assumes that the forecast uncertainty is constant over time. If the uncertainty varies over time (e.g., during ENSO events), then the mean-square error may under- or overestimate the uncertainty of a particular forecast. Ideally, the spread of the forecast ensemble should provide a measure of the uncertainty in a particular forecast. However, the ensemble spread may also be systematically biased. Linear scaling has also been applied to correct conditional biases in the ensemble spread (Weigel et al. 2009; Eade et al. 2014). Figure 2b demonstrates the effect of linear scaling on the overconfident forecasts in Fig. 2a. Unconditional biases in the ensemble spread have also been estimated on shorter time scales (e.g., Gneiting et al. 2005).

Adjustment of each type of bias requires the estimation of at least one parameter. Both historical forecasts and corresponding observations are required in order to estimate the required adjustments based on past performance. Measurement error in the observations is an additional source of uncertainty when estimating the parameters of each adjustment (Weijs and van de Giesen 2011). The short length of the observational record and the computational expense of large hindcast experiments limit the amount of data available for training recalibration methods for seasonal-to-decadal forecasts. Fewer than 30 years of data (i.e., fewer than 30 forecasts) are commonly used (e.g., Arribas et al. 2011). If too many parameters are estimated using too few data, then the data can be overfitted. The fitted parameters may predict the noise in the training data, rather than any systematic relationship between the forecasts and the observations. Recalibrating new forecasts using those parameters may actually decrease the skill of the forecasts.

On the other hand, training on too much data may be as bad as training on too little. Much predictability on seasonal-to-decadal time scales is thought to arise from slowly varying processes in the oceans. The oceans experience long-term variability on time scales from a few months (the seasonal cycle) to a few years (e.g., ENSO; Trenberth et al. 2007, section 3.6.2), or a few decades [e.g., the Atlantic multidecadal oscillation (AMO); Trenberth et al. 2007, section 3.6.6]. Multidecadal fluctuations may give rise to long-term but reversible trends. Predictability itself may also vary over time (Goddard and Dilley 2005). Many seasonal forecasting centers now update their recalibration parameters before each forecast, by retraining on only the most recent p years; examples include ECMWF System 4 (Molteni et al. 2011) and the Met Office Global Seasonal Forecast System, version 5 (GloSea5; MacLachlan et al. 2015). This approach has the advantage that the recalibration will adapt to any changes in long-term trends, time-varying predictability, the observing network, etc.

This raises two key questions: What is the optimal training period for a particular recalibration method? What is the optimal recalibration method given a fixed amount of training data? The aim of this study is to provide a flexible methodology to address these questions, and to draw general conclusions based on analysis of existing forecast systems and hindcast experiments where possible.

When evaluating the performance of a recalibration method, the data should be split ideally into a training set and an evaluation set (i.e., out-of-sample evaluation). If recalibrated forecasts are evaluated on the same data that the recalibration was trained on, then performance may be overestimated (Wilks 2011, section 6.4.4). A large evaluation set is required to reliably establish the expected performance of the recalibrated forecasts. Fixing the size of the evaluation set determines the maximum training period that can be evaluated given the available observation and forecast data. Out-of-sample evaluation of a sufficiently large sample of forecasts may not be possible given the data limitations noted above.

Cross validation is often used when out-of-sample evaluation is not possible. Cross validation involves holding back one or more forecasts to form a validation set; the recalibration is then trained on the remaining forecasts, and applied to the validation data. When considering training periods shorter than the length of the available data, there are many possible training and evaluation sets to choose from. If training sets are selected at random, then the training data could come from times when the state of the climate system and the observing network were quite different from the evaluation set. Appropriate cross-validation methods are required to allow the comparison of different recalibration methods and training periods using limited data.

The remainder of this paper is structured as follows: Section 2 introduces a family of statistical recalibration methods that includes all of the common adjustments to the forecast mean and variance; section 3 describes a cross-validation procedure for comparing different recalibration methods and training periods; section 4 applies the methodology developed in the previous sections to data from the CMIP5 near-term experiments; and section 5 discusses how widely the conclusions of this analysis might be expected to apply and how the methodology can be extended.

2. Recalibrating probability forecasts

This study is concerned with techniques for obtaining calibrated probability forecasts of a climate variable y_τ observed at time τ, from an ensemble of m deterministic forecasts $x_\tau = \{x_{\tau 1}, x_{\tau 2}, \ldots, x_{\tau m}\}$ output by a climate model. It is assumed that T observation–ensemble pairs $\{y_\tau, x_\tau\}$ are available, relating to forecast times $\tau = 1, \ldots, T$. If multiple lead times are required, they must be recalibrated separately to account for the possibility of systematic errors evolving as the model integrations extend further into the future. The methodology developed here is applicable to both absolute value and anomaly forecasts.

a. A general recalibration framework

An extended version of the ensemble model output statistics (EMOS) technique introduced by Gneiting et al. (2005) is proposed. The observable climate y_τ at time τ is represented by a linear function of the forecast time τ and the mean of the forecast ensemble $\overline{x}_\tau = m^{-1}\sum_{i=1}^{m} x_{\tau i}$,

$$y_\tau = a + b\overline{x}_\tau + t\tau + \varepsilon_\tau, \tag{1}$$

where a, b, and t are parameters to be estimated and ε_τ is a random error with zero expectation. The constant offset a represents unconditional bias in the ensemble mean. The scale parameter b represents conditional bias in the ensemble mean. The second scale parameter t represents (linear) time-dependent bias in the ensemble mean. A perfect unbiased forecast model would have $a = 0$, $b = 1$, and $t = 0$.

The variance of the random error ε_τ is allowed to depend linearly on the sample variance of the forecast ensemble $s_\tau^2 = (m-1)^{-1}\sum_{i=1}^{m}(x_{\tau i} - \overline{x}_\tau)^2$,

$$\mathrm{var}(\varepsilon_\tau) = c^2 + d^2 s_\tau^2, \tag{2}$$

where c and d are also parameters to be estimated. The constant offset c^2 represents unconditional bias in the ensemble variance, and the scale parameter d^2 represents conditional bias in the ensemble variance. Writing the forecast variance in terms of c^2 and d^2 rather than c and d ensures that $\mathrm{var}(\varepsilon_\tau) > 0$ for all s_τ^2. The random errors ε_τ are assumed to be normally distributed and independent between forecast times τ. Therefore, the forecast distribution can be written compactly as

$$y_\tau \sim N(a + b\overline{x}_\tau + t\tau, c^2 + d^2 s_\tau^2). \tag{3}$$

It is common practice in seasonal-to-decadal forecasting to detrend both the observations y_τ and the ensemble means $\overline{x}_\tau$ prior to recalibration (e.g., Kharin et al. 2012; Smith et al. 2012; Eade et al. 2014). Detrending is usually performed by estimating linear trends in y_τ and $\overline{x}_\tau$ separately using ordinary least squares, and removing them. The method of ordinary least squares is equivalent to simple linear regression. In appendix A, it is shown that simple linear regression of y_τ and $\bar{x}_\tau$ on the forecast time τ separately is precisely equivalent to multiple linear regression of y_τ on both $\bar{x}_\tau$ and τ. This result extends naturally to the case where $d \neq 0$ (see appendix B for details). Therefore, Eq. (3) reduces the procedure of detrending and recalibrating seasonal-to-decadal forecasts to a single step.

b. *A family of recalibration methods*

Training the recalibration framework in Eq. (3) requires the estimation of the five parameters a, b, t, c, and d. A family of recalibration methods of systematically varying complexity can be constructed by imposing restrictions on individual parameters (e.g., $t = 0$ or $d = 0$). The notation abtcd refers to the most general method where all five parameters are free to vary. A number in place of any parameter (or parameters) indicates that the parameter is fixed at that value; for example, ab0cd indicates that $t = 0$ (i.e., EMOS).

The family of recalibration methods described by Eq. (3) includes both the climatological (a00c0) and trend (a0tc0) forecasts. These are classed as statistical forecasts, since the forecast mean and variance are both independent of the forecast ensemble (i.e., $b = d = 0$). If the ensemble is not informative for the forecast mean, then it does not seem reasonable to assume that it is informative for the forecast variance. Therefore, methods where the forecast variance depends on the ensemble ($d \neq 0$), but the forecast mean does not ($b = 0$), are not considered in this study.

Neglecting time-dependent biases for a moment (i.e., let $t = 0$), the selected family contains four methods of estimating the forecast mean that do depend on the ensemble mean: the raw ensemble mean (010), the scaled ensemble mean (0b0), unconditional bias adjustment (a10), and conditional bias adjustment (ab0). If an adjustment for time-dependent biases is included (i.e., $t \neq 0$), then there are four additional methods of estimating the forecast mean (01t, 0bt, a1t, and abt). For each of the eight methods of estimating the forecast mean, there are five possible methods of estimating the forecast uncertainty: the mean-square error of the recalibrated forecast means (c0), the ensemble variance (01), the scaled ensemble variance (0d), the shifted ensemble variance (c1), and the shifted and scaled ensemble variance (cd). Therefore, Eq. (3) describes a family of 42 recalibration methods, including the two statistical forecasts.

c. *Mean-centered recalibration*

Methods of similar complexity may appear to perform similarly, particularly when applied to absolute value forecasts. For example, the black ellipse in Fig. 3 represents the joint distribution of observations and ensemble mean forecasts from a hypothetical model that has a small unconditional bias in the mean, but no unconditional or time-dependent biases ($a \neq 0$, $b = 1$, and $t = 0$). The theoretical regression lines corresponding to only unconditional bias correction (a10c0; Fig. 3, solid black line) and only conditional bias correction (0b0c0; Fig. 3, dashed black line) are almost coincident inside the forecast ellipse. Therefore, forecasts recalibrated using either method will exhibit very similar performance.

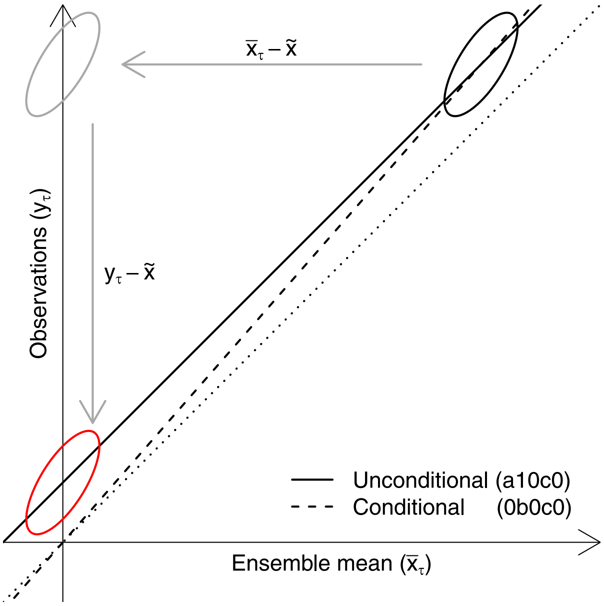

FIG. 3. Comparison of the mean-centered and general frameworks applied to absolute value forecasts. The black ellipse represents the joint distribution of observations and ensemble mean forecasts from a hypothetical model that has a small unconditional bias in the mean, but no conditional bias ($a \neq 0$, $b = 1$). The red ellipse represents the same forecasts after mean centering. The center of an ellipse corresponding to a "perfect" model with no unconditional bias would lie on the dotted line ($a = 0$, $b = 1$).

Different types of bias can be more easily distinguished by mean-centering the ensemble means $\bar{x}_\tau$ and forecast times τ. The mean-centered form of the general framework in Eq. (3) is given by

$$y_\tau \sim N[\tilde{x} + a + b(\bar{x}_\tau - \tilde{x}) + t(\tau - \tilde{\tau}), c^2 + d^2 s_\tau^2], \quad (4)$$

with

$$\tilde{x} = \frac{\sum_{\tau\in\gamma} w_\tau \bar{x}_\tau}{\sum_{\tau\in\gamma} w_\tau} \quad \text{and} \quad \tilde{\tau} = \frac{\sum_{\tau\in\gamma} w_\tau \tau}{\sum_{\tau\in\gamma} w_\tau},$$

where $\boldsymbol{\gamma} = \{\tau_1, \ldots, \tau_p\}$ is the set of p forecast times in the training set and

TABLE 1. Comparison of the complexity of existing forecast recalibration methods.

Complexity	Notation	Parameters	Method
0	01001	—	Raw ensemble forecast (Wilks 2002).
1	a1001	(a)	Mean bias adjustment.
2	a00c0	(a, c)	Climatological forecast (Stockdale 1997).
3	a0tc0	(a, t, c)	Trend forecast.
	a1tc0	(a, t, c)	Linear trend adjustment (Kharin et al. 2012).
	ab0c0	(a, b, c)	MOS (Glahn and Lowry 1972).
	ab00d	(a, b, d)	Ratio of predictable components (Kharin and Zwiers 2003; Eade et al. 2014).
4	ab0cd	(a, b, c, d)	EMOS (Gneiting et al. 2005).
	abtc0	(a, b, t, c)	MOS with linear trend adjustment.
	abt0d	(a, b, t, d)	Ratio of predictable components after detrending (Eade et al. 2014).
5	abtcd	(a, b, t, c, d)	EMOS with linear trend adjustment.

$$w_\tau^{-1} = \text{var}(\varepsilon_\tau) = c^2 + d^2 s_\tau^2. \tag{5}$$

The weighted means $\tilde{x}$ and $\tilde{\tau}$ simplify to unweighted means when $d = 0$.

The red ellipse in Fig. 3 represents the same forecasts after mean centering. The theoretical regression line corresponding to only unconditional bias correction is still the solid black line. However, the theoretical regression line corresponding to only conditional bias correction is now the dotted line ($a = 0, b = 1$). This line does not intersect the forecast ellipse at all. Therefore, the performance of the conditionally recalibrated forecasts will be very poor, as expected. Similar problems occur when trying to distinguish between two-parameter mean adjustments (i.e., ab0, 0bt, and a0t) estimated using the general framework.

The mean-centered framework also simplifies the interpretation of the recalibration parameters. It can be shown that the maximum likelihood estimate of a always has the form

$$\hat{a} = \tilde{y} - \tilde{x},$$

where $\tilde{y}$ is defined analogously to $\tilde{x}$ and $\tilde{\tau}$. Therefore, $\hat{a}$ can always be interpreted as an estimate of the expected bias between the forecasts and the observations.

d. Parameter estimation

The parameters of either the general or mean-centered families can be estimated by standard maximum likelihood methods. When $c = 0$ or $d = 0$, analytic expressions are available for the parameter estimates. Analytic expressions for the maximum likelihood estimates are not available when the forecast uncertainty is modeled by the shifted (c1) or shifted and scaled ensemble variance (cd). Therefore, parameter estimates must be obtained by numerical maximization of the likelihood. Full details are given in appendix B.

The scale parameter b is related to the correlation between the forecasts and observations. Negative correlations are difficult to interpret without a detailed understanding of the underlying physical processes. In this study, if a negative estimate of b was obtained analytically or numerically, then the parameters were reestimated with the additional constraint that $b = 0$.

e. Comparison to existing methods

Table 1 includes a number of examples of existing recalibration methods that can be considered special cases of the general framework presented here. In particular, when $t = 0$, Eq. (3) is equivalent to the EMOS method of Gneiting et al. (2005). Similarly, when $d = 0$, EMOS is equivalent to the traditional method of model output statistics (MOS; Glahn and Lowry 1972). Normal distributions have also previously been fitted to the raw ensemble forecast (i.e., 01001; Wilks 2002). The climatological forecast (a00c0) is often used as a benchmark for forecast skill (Stockdale 1997).

Kharin et al. (2012) performed a deterministic recalibration by estimating an unconditional adjustment after linearly detrending the observations y_τ and the ensemble means $\overline{x}_\tau$ separately by ordinary least squares. The observed trend was then added back to the detrended ensemble means to obtain recalibrated forecast means. As discussed in section 2a, this is equivalent to multiple linear regression of y_τ on both $\overline{x}_\tau$ and τ. Therefore, the method of Kharin et al. (2012) is equivalent to a1tc0.

The ratio of predictable components (RPC) method of Eade et al. (2014) involves a conditional adjustment of the ensemble mean after first removing the mean bias compared to the observations. The RPC estimate of the linear scaling parameter [Eq. (2); Eade et al. 2014] is equal to the maximum likelihood estimate from simple linear regression (Krzanowski 1998). Therefore, RPC correction of the ensemble mean is equivalent to ab0c0. RPC correction differs from the methodology presented here by issuing ensemble forecasts rather than probability forecasts. The individual ensemble members are rescaled about their recalibrated means, subject to the constraint that the total variance in the observations y_τ is equal to the sum of the variance explained by the recalibrated ensemble means and the variance in the individual members $x_{\tau i}$. Thus RPC correction can be seen as a hybrid between mean-square error (ab0c0) and scaled variance (ab00d) methods. The recalibrated forecast means will be equal to those from ab0c0. The forecast variances will be proportional to those estimated by ab00d, but not equal. An optional step in the RPC methodology is to first detrend both the observations y_τ and the ensemble means $\overline{x}_\tau$. In that case, RPC correction can be considered a hybrid of abtc0 and abt0d.

Coelho et al. (2004) also suggest an ab00d-type recalibration, but use a Bayesian approach to combine the ensemble forecast with a statistical forecast. The resulting posterior forecast distribution has a form similar to an ab0cd (EMOS) recalibration.

3. Comparing recalibration methods

The relative performance of different recalibration methods and training periods must be compared fairly in order to determine the most appropriate method and period for a particular forecast model and variable. Forecast performance is usually quantified using scoring rules. A score is assigned to each forecast by comparing it to a verifying observation according to the scoring rule. Proper scores have the desirable property that the forecaster cannot improve their expected score by hedging their forecasts (Jolliffe and Stephenson 2011, chapter 3). Examples of proper scores for probability forecasts include the ignorance score

$$S(f, y) = -\log f(y), \tag{6}$$

where f is the forecast density and y is the verifying observation, and the continuous ranked probability score (CRPS)

$$S(F, y) = \int_{-\infty}^{+\infty} [F(u) - H(u - y)]^2\, du, \tag{7}$$

where $F(u) = \int_{z \le u} f(z)\, dz$ is the cumulative forecast distribution and $H(u - y)$ is the Heaviside function (Matheson and Winkler 1976). Both are negatively oriented scores (i.e., lower values indicate better performance). Scores must be averaged over many forecasts in order to estimate the expected performance of a forecast system.

A structured cross-validation procedure is proposed to allow the comparison of training periods of different lengths given limited data. The cross-validated score at time τ is defined as the average score obtained from recalibrating the forecast for time τ using all possible continuous sets of $p + 1$ hindcasts that contain τ, but with τ omitted from each set (i.e., trained on p years of data). Let $F_{\tau,j}$ be the forecast distribution [Eq. (4)] for time τ, derived from a recalibration method trained on the $p + 1$ hindcasts beginning at time j, but excluding the forecast time τ. The score at time τ is defined as

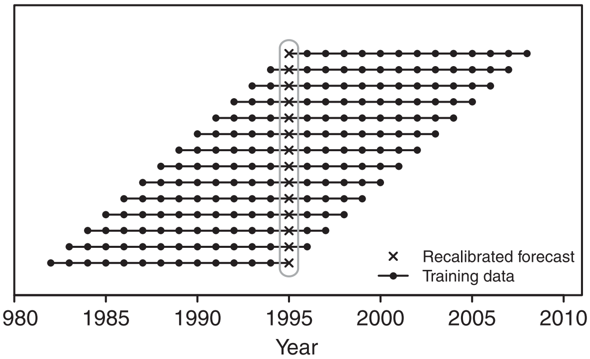

FIG. 4. Schematic representation of the proposed cross-validation procedure for a forecast of the period December–February in 1995 with a lead time of one month. The training period is 13 years (p = 13) and 30 years of hindcasts and corresponding observations are available (T = 30). Hindcasts are started every 12 months. Gray lines around forecasts indicate averaging over scores.

$$S(F_\tau, y_\tau) = \frac{1}{U(\tau) - L(\tau) + 1} \sum_{j=L(\tau)}^{U(\tau)} S(F_{\tau,j}, y_\tau), \tag{8}$$

where $F_\tau = \{F_{\tau,j} | L(\tau) \le j \le U(\tau)\}$ is the set of cross-validated forecast distributions $F_{\tau,j}$ for time τ, determined by the limits

$$L(\tau) = \max(1, \tau - p) \quad \text{and} \tag{9a}$$

$$U(\tau) = \min(\tau, T - p). \tag{9b}$$

This definition ensures that the training data are always contemporary to the forecast being evaluated. The procedure is illustrated in Fig. 4 for 1995, based on a training period of 13 years (p = 13), using 30 years of hindcast and observation data from 1981 to 2010 (T = 30). The score assigned to the forecast for 1995 is the average of 14 scores (Fig. 4, gray box). The time-varying bounds $L(\tau)$ and $U(\tau)$ fix the number of scores to be averaged over at each time τ.

Using the definition above, a cross-validated score can be computed for any time $1 \le \tau \le T$ for any length of training period $\pi < p < T$, where π is the number of mean parameters (a, b, t) to be estimated. An average cross-validated score can then be taken over any evaluation period required (i.e., any set of times τ). The evaluation period should be as long as possible in order to evaluate performance during different regimes and states of the climate system. Using the methodology proposed here, performance can be evaluated over all available forecast times.

This proposed cross validation is immediately applicable to seasonal forecasts where the successive forecasts do not overlap. The procedure is easily adapted to the case of overlapping forecasts (e.g., decadal forecasts) by excluding from each training set a block of one or more forecasts before or after the forecast being recalibrated. The same block adaptation can also be used to offset any other autocorrelation between successive forecasts.

4. Results from the CMIP5 near-term experiments

The methodology developed in sections 2 and 3 was applied to forecasts of monthly mean near-surface (2 m) temperature from the extended suite of CMIP5 near-term experiments (Taylor et al. 2012). The extended design consisted of 10 ensemble members initialized every year between 1960 and 2010. Only the CCCma CanCM4, Met Office Hadley Centre HadCM3, and NOAA/GFDL GFDL CM2.1 models completed the extended set of core experiments. The CanCM4 and GFDL CM2.1 models use full-field initialization. The HadCM3 model uses anomaly initialization by default. Similar results were obtained from the analysis of each model. Therefore, only results from the Canadian Centre for Climate Modelling and Analysis (CCCma) CanCM4 model are presented here (Merryfield et al. 2013). The CanCM4 hindcasts were initialized on 1 January each year. The ensemble members varied only in their initial conditions.

Forecast performance was evaluated for the mean temperature averaged over months 1 through 12 (January–December) following each start date. This forecast period was chosen as a compromise between the short lead times of seasonal forecasts and the longer averaging periods of decadal forecasts. It is also one of the evaluation periods recommended by Goddard et al. (2013). Forecasts were verified against observations from the Hadley Centre/Climatic Research Unit Temperature, version 3v (HadCRUT3v), dataset (Brohan et al. 2006). The HadCRUT3v data consist of anomalies relative to the period 1961–90. The HadCRUT3v anomalies were added to the climatology produced by Jones et al. (1999) to obtain absolute temperatures for comparison with the CanCM4 hindcasts. Before recalibration, the hindcasts were bilinearly interpolated to the latitude–longitude grid of the observations (5° × 5°). Only grid boxes with no missing monthly observations over the whole period 1961–2010 were included.

Training periods of 9, 13, 17, 21, 25, 29, 33, 37, 41, 45, and 49 yr were considered. All 42 recalibration methods belonging to the family described in section 2 were evaluated for all 11 training periods. The mean-centered recalibration framework (section 2c) was used throughout. Forecast performance was evaluated using the CRPS [Eq. (7)]. Scores were averaged over the whole period 1961–2010 using the cross-validation methodology described in section 3.

Current practice is to apply the same recalibration method and training period at all grid points, but to estimate the recalibration parameters separately for each grid point (e.g., Goddard et al. 2013; Molteni et al. 2011; MacLachlan et al. 2015). This practice is also adopted here. The average area-weighted score over all grid boxes with no missing observations is used as a summary measure to compare the performance of the different recalibration methods and training periods.

a. What is the optimal estimate of the forecast uncertainty?

The relative performance of the five methods of estimating the forecast uncertainty was similar for all eight methods of estimating the forecast means. Performance after conditional recalibration and trend adjustment of the forecast mean (abt) is compared in Fig. 5a. The mean-square error of the recalibrated ensemble mean (c0) consistently outperforms all other methods of estimating the forecast uncertainty when scores are averaged over all grid boxes. No contiguous spatial regions were found where the mean-square error was outperformed by any other method (not shown). There is no evidence of a useful skill–spread relationship between the ensemble variance and the forecast uncertainty. Only the shifted and scaled ensemble variance (cd) is able to approach the performance of the mean-square error as the training period increases and more data are available to estimate the additional variance parameter. However, the ensemble variance tends to be damped toward the mean-square error ($d \approx 0$, not shown). Therefore, no additional information is gained by including the ensemble variance, and the mean-square error estimate is preferred. The estimates of the scale parameter d from the scaled variance method (0d) suggest that the ensemble variance is too small in CanCM4 ($d > 1$, not shown). The shifted variance estimate (c1) is also able to inflate the forecast uncertainty compared to the raw ensemble variance ($c > 0$). However, there is less variation in the forecast uncertainty when the scaled variance method is used. Therefore, the

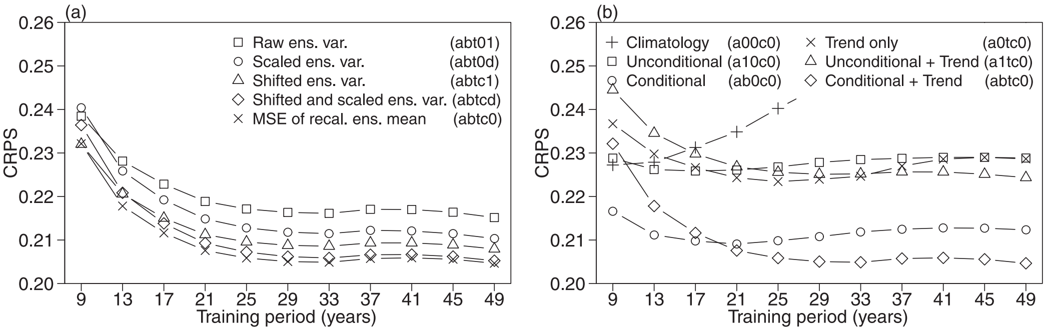

FIG. 5. Globally averaged scores comparing (a) methods of recalibrating the forecast uncertainty and (b) methods of recalibrating the forecast mean, computed using cross validation over 1961–2010.

shifted variance is closer in performance to the mean-square error.

b. What is the optimal estimate of the forecast mean?

The performance of the four methods of estimating the forecast mean that include unconditional bias adjustment is compared using the mean-square error estimate of the forecast uncertainty in Fig. 5b. The performance of methods not including unconditional bias adjustment was so poor that it would not be visible on the same scale (CRPS $\approx$ 0.65–0.67). Without unconditional bias adjustment, the forecast variance is inflated to account for the errors in the mean. This leads to forecasts with limited resolution and poor overall scores. This result emphasizes the importance of adjusting for unconditional biases as a minimum standard of recalibration.

However, conditional bias adjustment (ab0c0) clearly outperforms unconditional bias adjustment (a10c0) of the forecast mean. Trend-adjusted recalibration methods match or exceed the performance of their unadjusted equivalents when the training period is greater than 20 years. Similarly, the trend forecast (a0tc0) is able to outperform the climatological forecast (a00c0) if more than 15 years of observations are available.

c. Why is conditional bias adjustment important?

Figure 6b shows that conditional bias adjustment (ab0c0) consistently damps the ensemble mean toward the climatology ($b < 1$). Conditional bias adjustment with trend adjustment (abtc0) exhibits almost identical behavior (not shown). This agrees with previous studies that found that the ensemble mean may exhibit too much variability compared to the observations at the gridbox level (e.g., Goddard et al. 2013; Eade et al. 2014). The relative skill of the conditional and unconditional recalibrations was quantified by the continuous ranked probability skill score (CRPSS; appendix C). In CanCM4, the regions of strongest damping ($b \approx 0$) coincide with some of the largest gains in skill by conditional (ab0c0) compared to unconditional (a10c0) recalibration (Fig. 6a). In these regions, linear scaling of the ensemble mean ($b \neq 1$) allows the flexibility to issue climatological forecasts (a00c0) where the model has limited skill otherwise. Scaling the ensemble mean does not improve the minimum achievable score over all grid boxes (Fig. 7a). Instead, the average score is improved compared to unconditional bias adjustment by decreasing the number of grid boxes with high (poor) scores. Trend-adjusted conditional bias adjustment (abtc0) shows similar improvements compared to trend-adjusted unconditional recalibration (a1tc0; Fig. 7a).

d. What is the effect of trend adjustment?

Conditional bias adjustment (ab0c0) makes clear improvements to both the average score (Fig. 5b) and the distribution of scores (Fig. 7b) compared to unconditional bias adjustment (a10c0) alone. In contrast, the improvement in average score due to trend adjustment is small when optimal training periods are compared (Fig. 5b). There is almost no visible difference in the score distributions of similar recalibrations with and without trend adjustment (Fig. 7a). A direct comparison shows that trend-adjusted forecasts actually outperform equivalent forecasts based on conditional recalibration alone at almost 75% of grid boxes (Fig. 7b).

Trend adjustment is particularly effective in the tropics and at high latitudes (Fig. 6c). The rain forests in northeast Amazonia and West Africa have experienced particularly strong warming trends in recent decades (Malhi and Wright 2004). In the Arctic, climate models are known to have strong biases in their reproduction of recent trends in sea ice extent (Wang and Overland 2009). However, observation uncertainty may also play an important role at high latitudes, because of limited data. Recent trends in monsoon onset and rainfall in India have been linked to natural variability on interdecadal time scales as well as anthropogenic climate change (Krishnamurthy and Goswami 2000; Goswami et al. 2010; Sinha et al. 2015).

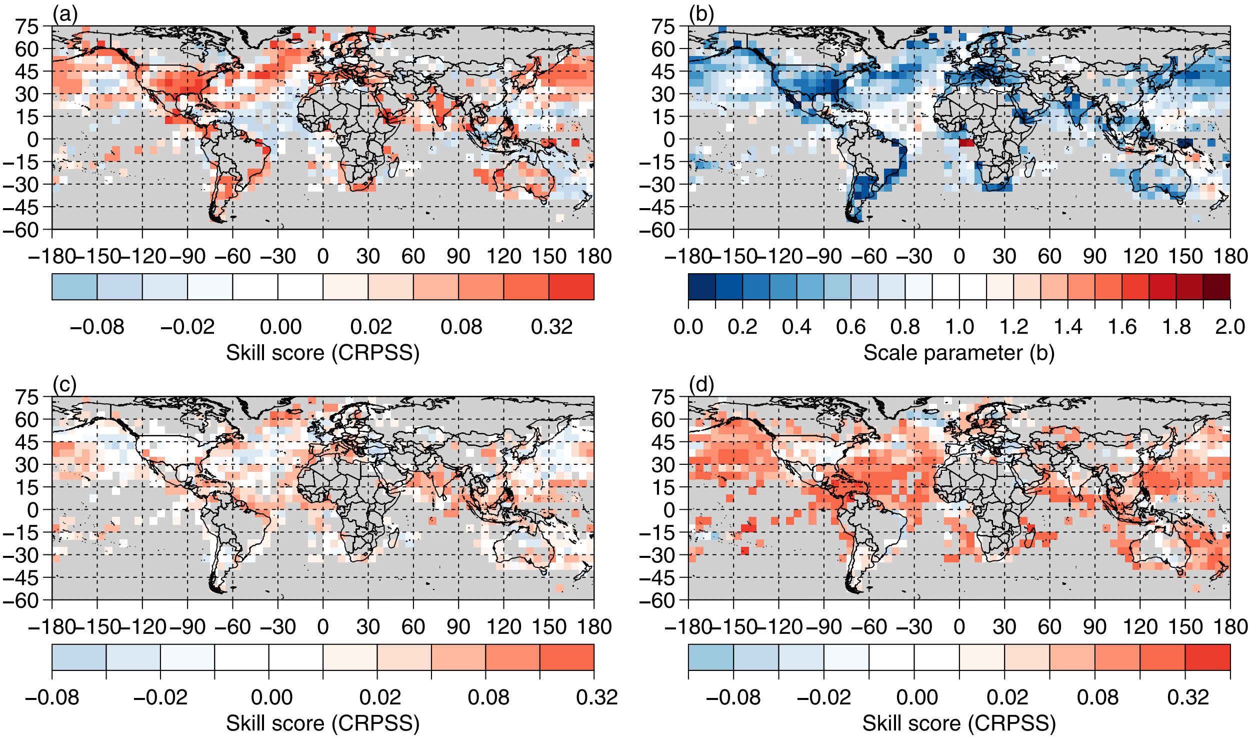

FIG. 6. (a) Skill of conditional bias adjustment (ab0c0) trained on 21 years of data compared to unconditional bias adjustment (a10c0) trained over 17 years; (b) average ensemble mean scale parameter b estimated by conditional bias adjustment (ab0c0) trained over 21 years; (c) skill of conditional bias adjustment with trend adjustment (abtc0) trained over 33 years, compared to conditional bias adjustment without trend adjustment (ab0c0) trained over 21 years; and (d) skill of conditional bias adjustment with trend adjustment (abtc0) trained over 33 years, compared to a trend forecast (a0tc0) trained over 25 years. All computed using cross validation over 1961–2010. Relative skill is compared using the CRPSS (appendix C).

Some of the largest improvements due to trend adjustment (e.g., Iceland, India, and the coast of Brazil; Fig. 6c) occur where the ensemble mean is strongly damped (Fig. 6b). When trend adjustment is included, the ensemble mean is damped toward the trend forecast (a0tc0) rather than a flat climatology. Trend adjustment in combination with conditional recalibration (abtc0) allows well calibrated trend forecasts (a0tc0) to be issued where both the forecast model and a flat climatology (a00c0) perform poorly.

In some locations, trend-adjusted forecasts can be outperformed by conditional or unconditional recalibration alone (Fig. 6c). However, the relative decrease in performance at these locations is generally small compared to the relative increase in performance elsewhere (Fig. 7c). Conditional bias adjustment with trend adjustment (abtc0) is able to achieve positive skill compared to an optimal trend forecast (a0tc0) at almost 85% of grid boxes (Fig. 6d).

e. *What is the optimal training period?*

The average score of recalibrated forecasts both with and without trend adjustment exhibits a nonlinear response to increasing training period (Fig. 5b). For trend-adjusted methods ($t \neq 0$), there is a locally optimum training period of around 29–33 yr, similar to the optimal trend forecast (25 yr). The optimal training period for methods without trend adjustment ($t = 0$) is shorter, around 17–21 yr, similar to the optimal climatological forecast (9–13 yr). In both cases, the average score degrades once the training period exceeds the local optimum, before improving again above 40 years. In terms of the gridbox-averaged score, the optimum training period for both unconditional and conditional bias adjustment with trend adjustment is actually 49 years.

Without trend adjustment, the optimum period tends to be either very long or very short (Figs. 8a,c). After trend adjustment, the number of grid boxes where very long training periods are preferred increases by almost 50% (Figs. 8b,d). Short training periods are only optimal at a small number of locations. However, training periods close to the optimal trend forecast (21–33 yr) are preferred at many grid boxes.

Increasing the training period of trend-adjusted conditional bias adjustment from the local optimum of 33 years

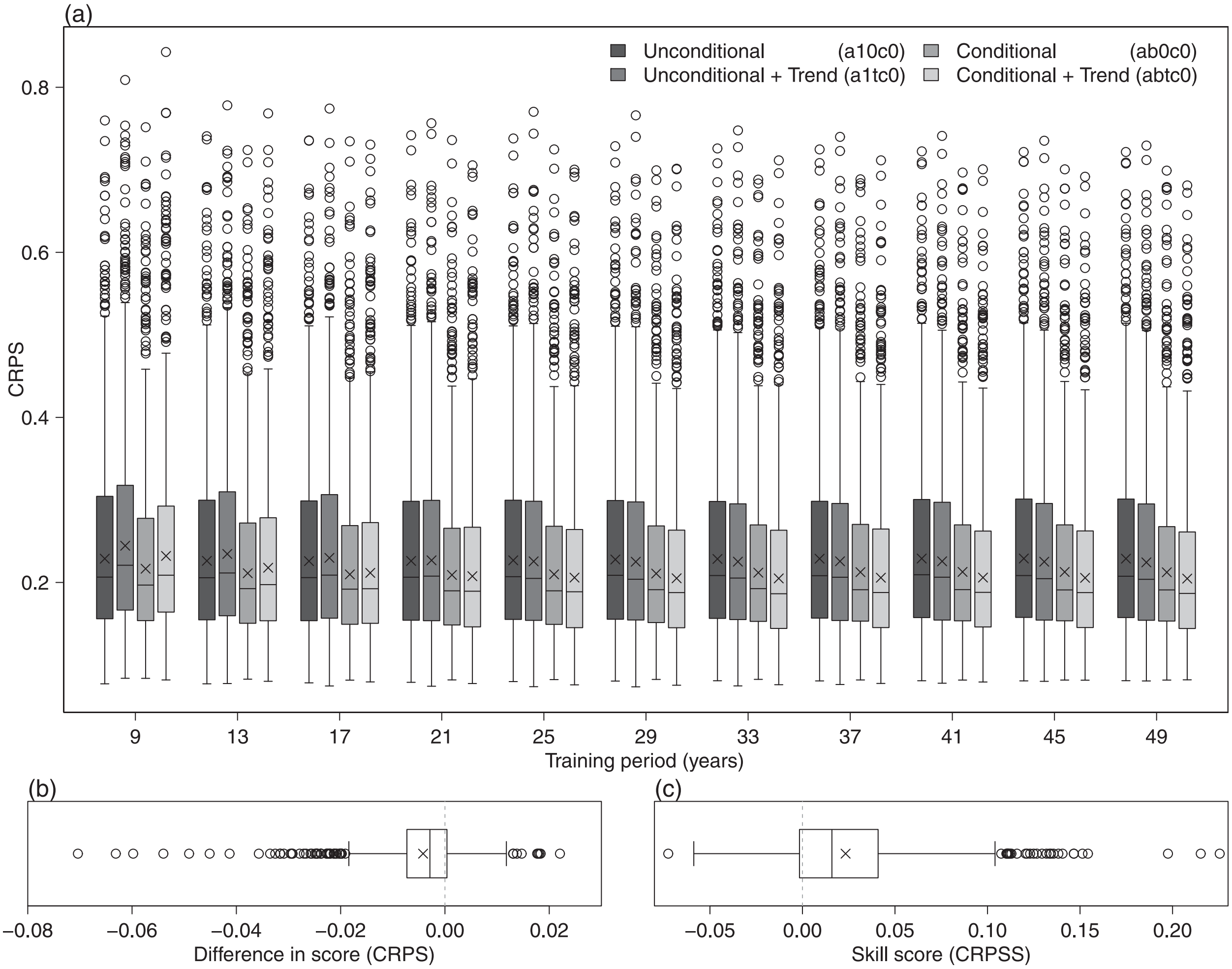

FIG. 7. Distribution of (a) gridbox scores using unconditional or conditional recalibration, with or without trend adjustment; (b) difference in gridbox scores between conditional recalibration with (abtc0, 33 yr) and without (ab0c0, 21 yr) trend adjustment; and (c) gridbox skill comparing conditional recalibration with (abtc0, 33 yr) and without (ab0c0, 21 yr) trend adjustment. All computed using cross validation over 1961–2010. The whiskers of the box-and-whisker plots indicate the lowest and highest data points within 1.5 times the interquartile range of the lower and upper quantiles respectively. Data points outside this range are marked with circles. Crosses indicate the area-weighted mean score (or skill score, as appropriate).

to the maximum of 49 years increases performance at more than 60% of grid boxes (Fig. 9a). However, performance decreases where shorter training periods are preferred. Performance near Iceland, the coast of Brazil, southern Europe and the Mediterranean region, India, and the Arabian Sea is particularly badly affected. These regions were previously identified as regions where trend adjustment increases performance (Fig. 6c), but where the ensemble mean is strongly damped toward the trend forecast (Fig. 6b). Acceptable performance in these regions is dependent on making well-calibrated statistical (trend) forecasts. Thus, performance is strongly affected by increasing the training period away from the optimal trend forecast. Small increases in performance at the majority of grid boxes are balanced by large decreases where the model has limited skill and well-calibrated trend forecasts are required (Fig. 9b). As a result, the area-averaged skill score actually favors the shorter training period.

The optimum training period is a compromise. Long training periods are optimal at many grid boxes without trend adjustment. However, performance at many other locations is optimized by a climatological forecast with a short training period (9–13 yr). The optimal training period without trend adjustment (17–21 yr) is a balance between these competing factors. Long training periods are optimal at even more locations when trend adjustment is applied. However, the increase in performance at the majority of locations is outweighed by strong reductions in performance where well-calibrated (25 yr)

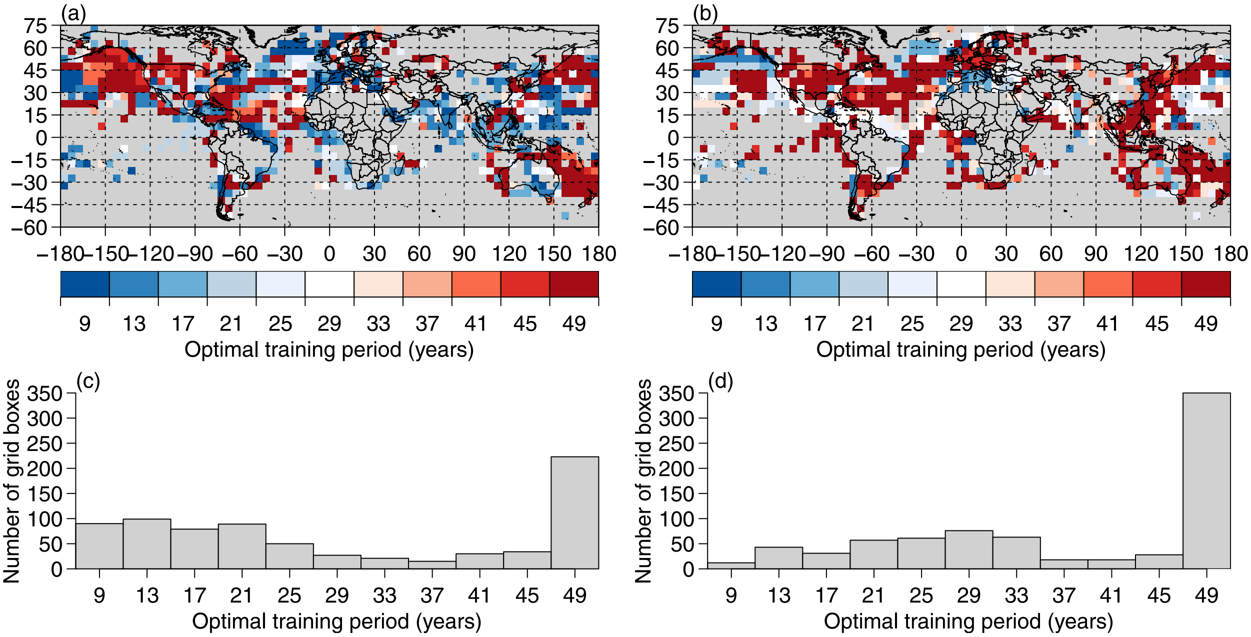

FIG. 8. The optimal training period (a),(c) before (ab0c0) and (b),(d) after (abtc0) trend adjustment. The optimal training period is based on the minimum score after conditional trend adjustment, computed using cross validation over 1961–2010.

trend forecasts are optimal. Again, the overall optimal training period (29–33 yr) strikes a balance between the two time scales. Higher overall performance may be possible using very long (>50 yr) training periods, but at the expense of very poor performance in a few locations.

f. Comparison with out-of-sample performance

Cross validation might be expected to overestimate forecast performance through the inclusion of training data that postdate the forecasts. Some studies have found that cross validation can actually underestimate forecast performance (Smith et al. 2013). Forecast performance estimated by cross validation was compared to out-of-sample performance using a shorter evaluation period. Out-of-sample forecasts for 1991–2010 were produced using rolling training periods of between 9 and 29 yr. Cross-validated average scores were produced for the same period [Eq. (8)]. The results of both analyses are qualitatively in good agreement. The mean-square error consistently outperforms other methods of estimating the forecast uncertainty analyses (not shown). Conditional bias adjustment consistently outperforms unconditional bias adjustment (Figs. 10a,b). Trend-adjusted methods outperform equivalent unadjusted methods when trained on more than 20 years of hindcasts.

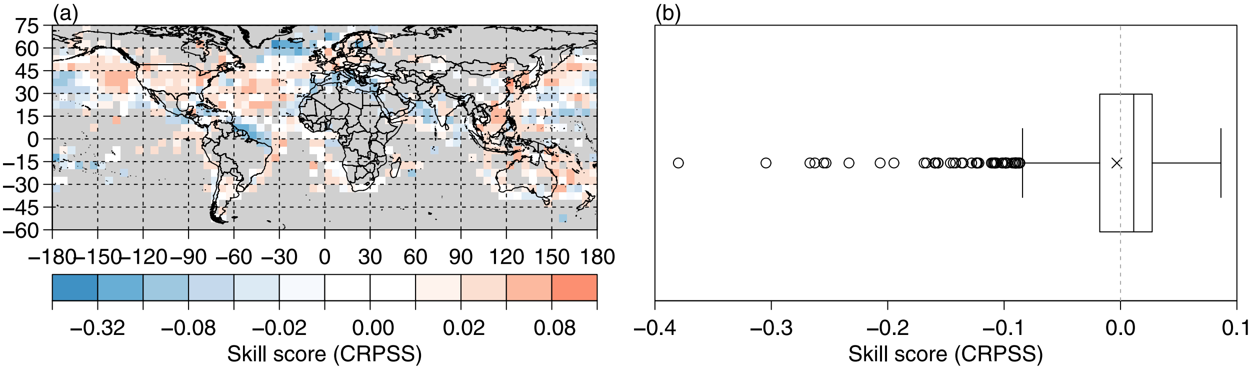

FIG. 9. (a) The spatial distribution and (b) box-and-whisker plot of the distribution of the skill of conditional recalibration with trend adjustment (abtc0) for 1961–2010 after training on 49 years of data compared to training on 33 years of data. The whiskers of (b) are computed as in Fig. 7.

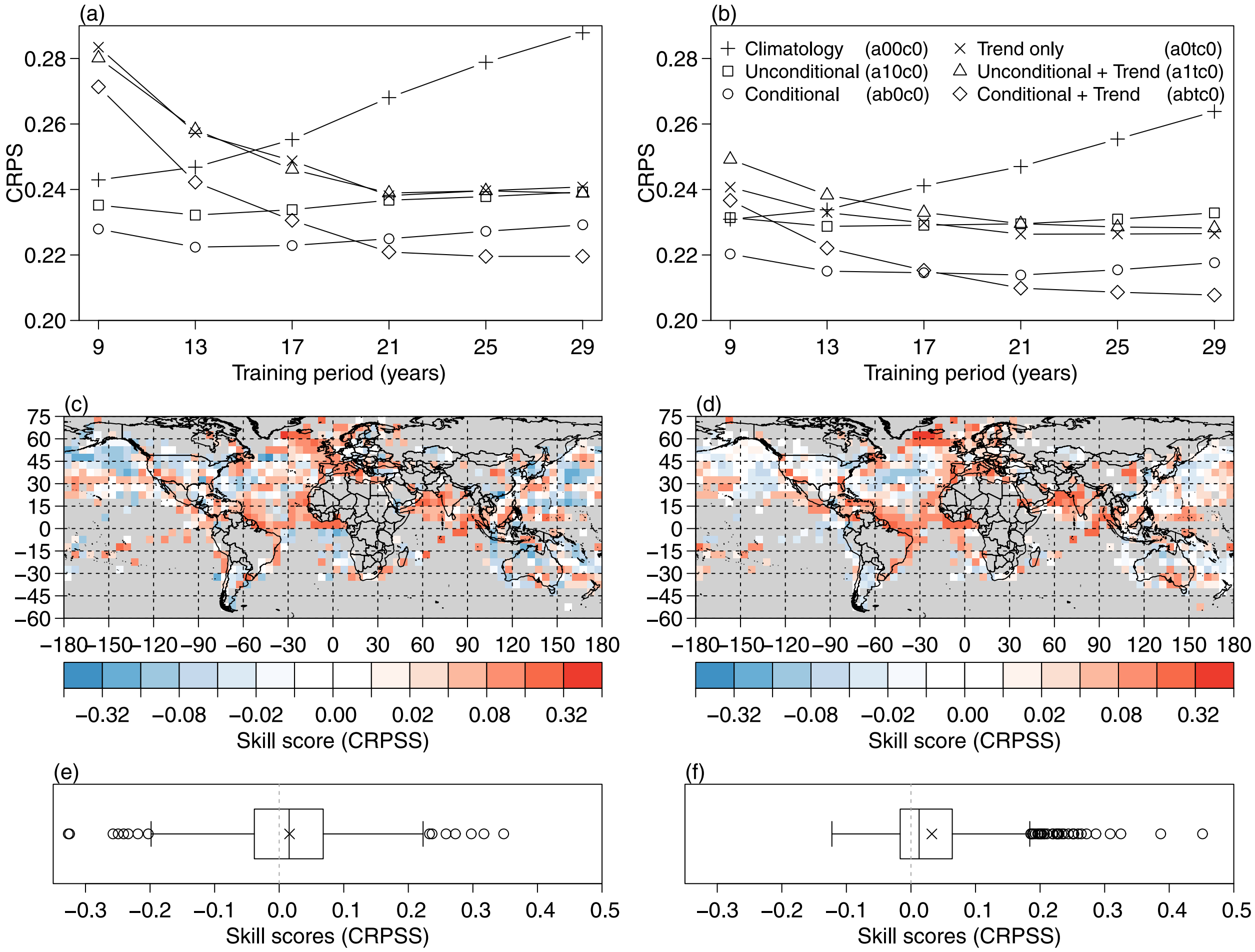

FIG. 10. Comparison of methods of recalibrating the forecast mean computed using (a) out-of-sample forecasts and (b) cross validation. Comparison of gridbox skill of conditional recalibration with (abtc0, 29 yr) and without (ab0c0, 13 yr) trend adjustment based on (c),(e) out-of-sample forecast and (d),(f) cross validation. All scores are averaged over 1991–2010. The whiskers of the box-and-whisker plots are computed as in Fig. 7. Crosses in (e),(f) indicate area-averaged skill.

Cross validation tends to overestimate the performance of all methods, particularly for short training periods (Figs. 10a,b). The optimal training period also tends to be slightly overestimated. Both overestimates are likely to be caused by the inclusion of training data that postdate the forecasts during cross validation. The spatial pattern of skill of trend-adjusted compared to unadjusted conditional bias adjustment is reproduced well by cross validation (Fig. 10c). However, the performance of trend adjustment tends to be overestimated in regions where it may not be optimal (e.g., the northeastern and northwestern Pacific). As a result, the overall advantage of trend adjustment (estimated by the area-averaged skill score) may be overestimated by the cross-validated results (Figs. 10e,f). Overall, the conclusions from the cross-validated analysis are consistent with the out-of-sample analysis. The conclusions of both analyses of 1991–2010 are also consistent with the cross-validated analysis of the full period 1961–2010.

g. *Selection bias*

By restricting the analysis to only grid boxes with no missing observations the study area is substantially restricted. This represents a potential source of bias in the results. To assess the robustness of the results to this selection bias, a similar analysis was performed after training and verifying using ERA-Interim (Dee et al. 2011) rather than HadCRUT3v. ERA-Interim is only available from 1979 onward. Therefore, the maximum training period is 30 years, and a cross-validated analysis is required.

The results are qualitatively consistent with those from the analysis using HadCRUT3v (Fig. 11a). Conditional bias adjustment consistently outperforms

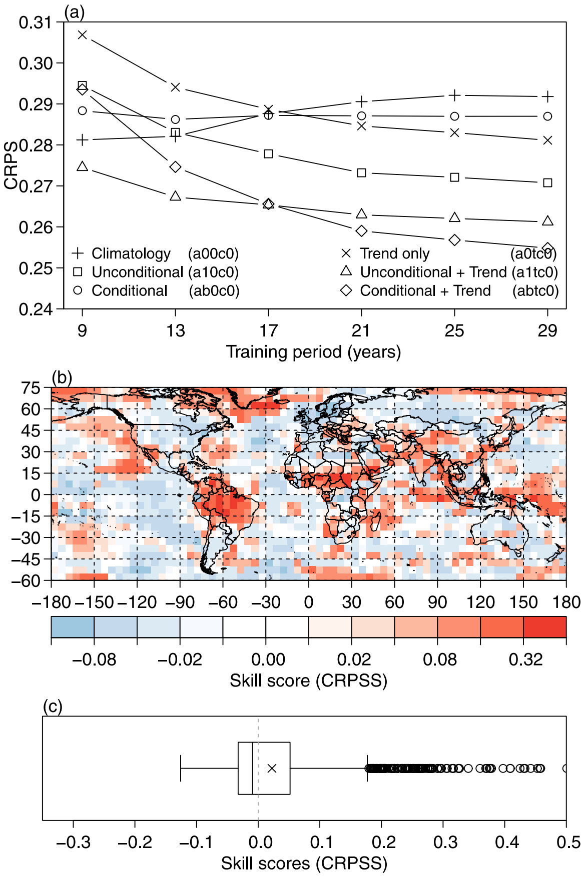

FIG. 11. (a) Comparison of methods of recalibrating the forecast mean and (b),(c) the distribution of gridbox skill after conditional bias adjustment with (abtc0, 29 yr) and without (ab0c0, 13 yr) trend adjustment, computed using cross validation over 1991–2010, recalibrated and verified relative to ERA-Interim. The whiskers of the box-and-whisker plot are computed as in Fig. 7.

unconditional bias adjustment. Trend-adjusted methods outperform their unadjusted equivalents when trained on more than 20 years of data. The optimal training period after trend adjustment is around 30 years. The main difference compared to the analysis based on HadCRUT3v in Fig. 10 is that trend-adjusted additive recalibration (a1tc0) is consistently outperformed by the trend forecast (a0tc0). Examination of the scale parameter *b* shows that the ensemble mean is strongly damped toward the climatological or trend forecast over large areas (not shown). Therefore, the overall poor performance of unconditional recalibration is unsurprising.

Trend-adjusted conditional bias adjustment (abtc0) outperforms unadjusted conditional bias adjustment (ab0c0) at only 45% of grid boxes (Fig. 11c). However, the relative decrease in performance where trend adjustment is not optimal is small compared to the relative increase elsewhere. Therefore, trend-adjusted conditional recalibration (abtc0) is still preferred overall.

The largest increases in performance due to trend adjustment occur over tropical landmasses and around the Arctic ice edge in ERA-Interim (Fig. 11b). Similar performance increases due to trend adjustment occur when trained and verified against HadCRUt3v (Fig. 10d). Although the overall results from ERA-Interim and HadCRUT3v are qualitatively consistent, there are also differences (e.g., southern Europe). It is important to remember that reanalysis data are not observations. If the reanalysis model and the forecast model possess similar deficiencies, then the forecasts may appear more skillful than if they had been verified against real observations. However, reanalysis data provide a valuable check for selection bias due to incomplete observation data.

5. Discussion

This study has developed a unified framework for the evaluation of statistical recalibration methods for seasonal-to-decadal probability forecasts. The process of detrending forecasts and observations to adjust for differences in linear time trends has been integrated into the recalibration procedure by exploiting results from multiple linear regression. The concept of a family of recalibration methods of systematically varying complexity has been introduced. A new cross-validation methodology has been developed to address the unique features of seasonal-to-decadal climate forecasts. Analysis of the CMIP5 hindcasts shows that conditional adjustment of the ensemble mean is required to obtain reliable forecasts where the model has limited skill. Conditional adjustment combined with trend adjustment allows the use of long training periods and further increases forecast skill.

The general approach developed here can be broken down into four stages. First, a family of recalibration methods is identified. Second, a range of training periods is selected. Third, an evaluation period is selected and cross validation applied to obtain recalibrated forecasts from each method and training period. Finally, scoring rules are applied to evaluate the performance of each recalibration method and training period.

The family of recalibration methods introduced here is very flexible, but not without limitations. The current framework is restricted to linear adjustments of the forecasts' mean, variance, and time trends. This assumption is likely to be overly simplistic, but trying to fit a more complicated model to so little data would risk overfitting. The assumption of normally distributed forecast errors will not be appropriate for all variables.

Other studies have extended the EMOS method to other error distributions (e.g., Scheuerer 2014; Baran and Nemoda 2016). The approach developed here to selecting an optimal method and training period can also be applied to those frameworks. The assumption of independent forecast errors may also be a limiting factor for some slowly evolving variables. This can be partially addressed by the use of blocking in the cross-validation step; however, further research is required to fully address this issue. The effect of uncertainty in the recalibration parameters is also routinely ignored. This issue has recently been addressed for regression frameworks, such as the one presented here, by Siegert et al. (2016). However, additional development is still required to account for observational uncertainty during recalibration.

The optimal recalibration method and training period will vary for different models, lead times, and forecast variables. However, the analysis of the CanCM4 hindcasts demonstrated some features that are expected to be present in other forecasts. Similar conclusions were also drawn from analysis of other CMIP5 models with varying initialization strategies (not shown). In particular, conditional bias adjustment allows the ensemble mean to be damped toward a suitable statistical forecast. At the gridbox level, there are always likely to be regions where a particular model has limited skill and might be outperformed by a well-calibrated statistical forecast. Therefore, the apparent relationship between the optimal training period and the length of the optimal statistical forecast is likely to be a general one. The strength of that relationship will depend on the proportion of grid boxes where the statistical forecast is optimal.

Trend adjustment also had a positive impact on forecast skill. In general, trend adjustment acts to lengthen the optimal training period by reducing time-dependent biases. Performance was improved both where little scaling of the ensemble mean was required, and where the ensemble mean was strongly damped. However, at locations where trend adjustment was not required or had no effect on the optimal training period, the forecast performance was reduced. So the overall increase in performance tended to be small. The tendency to issue strongly damped, trend-adjusted forecasts in regions experiencing rapid climate change emphasizes the need for seasonal-to-decadal forecast models to correctly reproduce climate trends.

In the analysis presented here, a single optimal recalibration method and training period were selected on the basis of the average area-weighted score over all grid boxes. Box-and-whisker plots of gridbox scores, score differences, and skill scores were also shown to be useful tools for analyzing the relative performance of different recalibration methods and training periods. For the CanCM4 hindcasts, training periods approaching 50 years were shown to have similar overall skill to training periods of around 30 years. Small increases in skill at the majority of grid boxes were balanced by large decreases in skill where trend forecasts were preferred. The benefits of long training periods appear to be limited compared to the additional computational expense. A more focused analysis might allow different recalibration methods or/and training periods to be selected in different regions.

The mean-square error of the recalibrated forecasts consistently outperformed other methods of estimating the forecast uncertainty in the analysis of CanCM4. It is possible that the ensemble size was too small to allow the reliable estimation of any skill–spread relationship in the ensemble variance. The effect of ensemble size shall be investigated further in Part II of this study. It is unlikely that the mean-square error is the optimal estimate of the forecast uncertainty for all models, variables, lead times and spatial scales. Also, it is important to test all combinations of appropriate mean and variance adjustments. In the analysis presented here, the relative performance of the five forecast variance estimates was similar across all of the forecast mean estimates. However, it is simple to construct cases where there would be strong interactions; for example, the raw ensemble variance might be an excellent estimate of the forecast uncertainty, but only after any biases in the mean have been removed.

The cross-validation methodology introduced here allows the comparison of different recalibration methods with varying training periods given limited data for any required validation period. The cross-validated analysis agreed qualitatively with the out-of-sample results. Cross validation tends to slightly overestimate the length of the optimal training period. However, the existence of an optimal training period less than the maximum training period was well captured, and the overestimation was small. The overall skill was also relatively insensitive to small changes in the training period when close to the optimum. Therefore, the cross-validation methodology presented here is recommended for the analysis of competing recalibration methods and training periods.

This study has focused on the common practice of recalibrating each grid box separately. However, this can lead to unexpected results such as the anomalous strong positive scale parameters estimated at a handful of grid boxes in Fig. 6b. Smoothing the data prior to recalibration may help to reduce the occurrence of such anomalies (Goddard et al. 2013). A more holistic approach would be to specify a covariance structure between the grid boxes and to estimate the recalibration parameters at each location simultaneously. This would lead to smoothly varying recalibration parameters that might be interpreted more easily. Ensemble copula coupling provides an intermediate approach that retains the spatial covariance structure in the forecasts, but does not require that the recalibration parameters vary smoothly (Schefzik et al. 2013).

In Part II of this study, the cross-validation methodology is extended to analyze the effects of ensemble size and hindcast frequency, and the optimal design of new forecast systems and hindcast experiments. In addition, the optimal recalibration methods and training periods of other forecast models, lead times, and averaging periods are investigated.

Acknowledgments. The authors were supported by a grant from the National Oceanic and Atmospheric Administration (NOAA) NA12OAR4310086. The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its subagencies.

APPENDIX A

Multiple Linear Regression

To see that Eq. (3) is equivalent to regression of the observations y_τ on the forecast means $\overline{x}_\tau$ after first regressing both on the time τ, write

$$\begin{aligned} y_\tau &= a_y + t_y\tau + \varepsilon_{y\tau}, \\ \overline{x}_\tau &= a_x + t_x\tau + \varepsilon_{x\tau}, \quad \text{and} \\ \varepsilon_{y\tau} &= b\varepsilon_{x\tau} + \varepsilon_\tau . \end{aligned}$$

Substitute first for $\varepsilon_{y\tau}$, and then for $\varepsilon_{x\tau}$ to obtain

$$\begin{aligned} y_\tau &= (a_y - ba_x) + b\overline{x}_\tau + (t_y - bt_x)\tau + \varepsilon_\tau \\ &= a + b\overline{x}_\tau + t\tau + \varepsilon_\tau , \end{aligned}$$

where $a = a_y - ba_x$ and $t = t_y - bt_x$.

APPENDIX B

Maximum Likelihood Estimation

The log–likelihood for the general framework in Eq. (3) is given by

$$\begin{aligned} \log\mathscr{L}(a,b,t,c,d;\mathbf{y}) = -\frac{p}{2}\log 2\pi + \frac{1}{2}\sum_{\tau\in\gamma}\log w_\tau \\ -\frac{1}{2}\sum_{\tau\in\gamma} w_\tau(y_\tau - \mu_\tau)^2 , \end{aligned}$$

with

$$\mu_\tau = a + b\overline{x}_\tau + t\tau \quad \text{and} \quad w_\tau^{-1} = c^2 + d^2 s_\tau^2 ,$$

where $\boldsymbol{\gamma} = \{\tau_1, \ldots, \tau_p\}$ is the set of forecast times in the training set, and p is the length of the training set.

Provided the weights w_τ are known up to a multiplicative constant, maximizing the log–likelihood is equivalent to minimizing the sum of squares

$$\log\mathscr{L} \propto (\mathbf{y} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{W}(\mathbf{y} - \boldsymbol{\mu}),$$

where $\mathbf{y}$ is the p vector of observations, $\boldsymbol{\mu}$ is the p vector of forecast means, and $\mathbf{W}$ is the diagonal $p \times p$ matrix with elements $W_{\tau\tau} = w_\tau$ and zero everywhere else. The mean vector can be written as $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta} = [a\, b\, t]^{\mathrm{T}}$ is the vector of mean parameters and $\mathbf{X}$ is the $p \times 3$ matrix with rows $[1\ x_\tau\ \tau]$ for τ in $1, \ldots, p$.

When $d = 0$ so that $w_\tau^{-1} = c$, the maximum likelihood estimates of the parameters $\boldsymbol{\beta}$ are equal to the ordinary least squares estimates given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

and

$$\hat{c}^2 = \frac{1}{p}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

Similarly, when $c = 0$ so that $w_\tau^{-1} = d^2 s_\tau^2 \propto s_\tau^2$, the maximum likelihood estimates of the parameters $\boldsymbol{\beta}$ are equal to the weighted least squares estimates given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{y}$$

and

$$\hat{d}^2 = \frac{1}{p}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{W}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

When the forecast uncertainty is modeled by the shifted (c1) or shifted and scaled ensemble variance (cd), the parameter estimates must be obtained by numerical maximization of the likelihood. This procedure can be simplified by noting that the log–likelihood depends on the variance parameters c and d only through the weights w_τ. So for given values of c and d, the maximum likelihood estimates of the mean parameters a, b, and t are obtained by weighted least squares. Given initial guesses for c and d, estimation proceeds by alternating between weighted least squares steps for the mean parameters, and Newton–Raphson (or similar) steps for the variance parameters.

This simplification reduces the dimension of the numerical optimization problem from five parameters to just two. The simplified optimization is more robust than maximizing the full five-dimensional log–likelihood, particularly when $b = 0$ where the solution was found to be very sensitive to the initial guesses for the parameters. This iterative procedure is equivalent to maximizing the profile log–likelihood for c and d (Garthwaite et al. 2002). Therefore, the parameter estimates will be identical to those obtained by maximizing the full five-dimensional likelihood.

The mean-centered framework in Eq. (4) can be fitted using the same procedure, but substituting $\mathbf{y} = \mathbf{y}'$ and $\mathbf{X} = \mathbf{X}'$, where $\mathbf{y}' = \mathbf{y} - \tilde{\mathbf{x}}$, with $\tilde{\mathbf{x}}$ the p vector of weighted ensemble means, and $\mathbf{X}'$ is the $p \times 3$ matrix with rows $[1 \quad \overline{x}_\tau - \tilde{x} \quad \tau - \tilde{\tau}]$. Note that the weighted means $\tilde{x}$ and $\tilde{\tau}$ must be recomputed after each Newton–Raphson step, since the weights depend on c and d.

Gneiting et al. (2005) suggested minimizing the continuous ranked probability score as an alternative to maximizing the likelihood. However, Williams et al. (2014) found little difference in the performance of forecasts estimated by either procedure, but noted that the computational cost of minimum CRPS estimation was much greater than that of maximum likelihood.

APPENDIX C

The Continuous Ranked Probability Skill Score

The continuous ranked probability skill score (CRPSS) can be used to compare the relative performance of two recalibration methods, or a set of recalibrated forecasts and a set of reference forecasts. If $\mathrm{CRPS}_{\mathrm{for}}$ and $\mathrm{CRPS}_{\mathrm{ref}}$ are the average scores of the recalibrated forecasts and reference forecasts respectively, then the CRPSS is given by

$$\mathrm{CRPSS} = \frac{\mathrm{CRPS}_{\mathrm{ref}} - \mathrm{CRPS}_{\mathrm{for}}}{\mathrm{CRPS}_{\mathrm{ref}}}.$$

The skill score provides a direct comparison of the performance of two sets of forecasts on a dimensionless scale. If there is no difference in average performance, then the skill score should equal zero. The skill score is bounded above at one for a perfect forecast, but has no fixed lower bound.

REFERENCES

Arribas, A., and Coauthors, 2011: The GloSea4 ensemble prediction system for seasonal forecasting. *Mon. Wea. Rev.*, **139**, 1891–1910, doi:10.1175/2010MWR3615.1.

Baran, S., and D. Nemoda, 2016: Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics*, **27**, 280–292, doi:10.1002/env.2391.

Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *J. Geophys. Res.*, **111**, D12106, doi:10.1029/2005JD006548.

Coelho, C. A. S., S. Pezzulli, M. Balmaseda, F. J. Doblas-Reyes, and D. B. Stephenson, 2004: Forecast calibration and combination: A simple Bayesian approach for ENSO. *J. Climate*, **17**, 1504–1516, doi:10.1175/1520-0442(2004)017<1504:FCACAS>2.0.CO;2.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, doi:10.1002/qj.828.

Eade, R., D. Smith, A. Scaife, E. Wallace, N. Dunstone, L. Hermanson, and N. Robinson, 2014: Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophys. Res. Lett.*, **41**, 5620–5628, doi:10.1002/2014GL061146.

Garthwaite, P., I. Jolliffe, and B. Jones, 2002: *Statistical Inference*. 2nd ed. Oxford University Press, 342 pp.

Glahn, B., M. Peroutka, J. Wiedenfeld, J. Wagner, G. Zylstra, B. Schuknecht, and B. Jackson, 2009: MOS uncertainty estimates in an ensemble framework. *Mon. Wea. Rev.*, **137**, 246–268, doi:10.1175/2008MWR2569.1.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, doi:10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, doi:10.1175/MWR2904.1.

Goddard, L., and M. Dilley, 2005: El Niño: Catastrophe or opportunity. *J. Climate*, **18**, 651–665, doi:10.1175/JCLI-3277.1.

——, and Coauthors, 2013: A verification framework for interannual-to-decadal predictions experiments. *Climate Dyn.*, **40**, 245–272, doi:10.1007/s00382-012-1481-2.

Goswami, B. N., J. R. Kulkarni, V. R. Mujumdar, and R. Chattopadhyay, 2010: On factors responsible for recent secular trend in the onset phase of monsoon intraseasonal oscillations. *Int. J. Climatol.*, **30**, 2240–2246, doi:10.1002/joc.2041.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447, doi:10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2.

ICPO, 2011: Decadal and bias correction for decadal climate predictions. International CLIVAR Project Office Tech. Rep. 150, 3 pp. [Available online at http://www.wcrp-climate.org/decadal/references/DCPP_Bias_Correction.pdf.]

Jolliffe, I. T., and D. B. Stephenson, Eds., 2011: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. 2nd ed. Wiley, 292 pp.

Jones, P. D., M. New, D. E. Parker, S. Martin, and I. G. Rigor, 1999: Surface air temperature and its changes over the past 150 years. *Rev. Geophys.*, **37**, 173–199, doi:10.1029/1999RG900002.

Kharin, V. V., and F. W. Zwiers, 2003: Improved seasonal probability forecasts. *J. Climate*, **16**, 1684–1701, doi:10.1175/1520-0442(2003)016<1684:ISPF>2.0.CO;2.

——, G. J. Boer, W. J. Merryfield, J. F. Scinocca, and W.-S. Lee, 2012: Statistical adjustment of decadal predictions in a changing climate. *Geophys. Res. Lett.*, **39**, L19705, doi:10.1029/2012GL052647.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2010: Challenges in combining projections from multiple climate models. *J. Climate*, **23**, 2739–2758, doi:10.1175/2009JCLI3361.1.

Krishnamurthy, V., and B. N. Goswami, 2000: Indian monsoon–ENSO relationship on interdecadal timescale. *J. Climate*, **13**, 579–595, doi:10.1175/1520-0442(2000)013<0579:IMEROI>2.0.CO;2.

Krzanowski, W. J., 1998: *An Introduction to Statistical Modelling.* John Wiley & Sons, 264 pp.

MacLachlan, C., and Coauthors, 2015: Global Seasonal Forecast System version 5 (GloSea5): A high-resolution seasonal forecast system. *Quart. J. Roy. Meteor. Soc.*, **141**, 1072–1084, doi:10.1002/qj.2396.

Malhi, Y., and J. Wright, 2004: Spatial patterns and recent trends in the climate of tropical rainforest regions. *Philos. Trans. Roy. Soc. London*, **359B**, 311–329, doi:10.1098/rstb.2003.1433.

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Manage. Sci.*, **22**, 1087–1096, doi:10.1287/mnsc.22.10.1087.

Merryfield, W. J., and Coauthors, 2013: The Canadian Seasonal to Interannual Prediction System. Part I: Models and initialization. *Mon. Wea. Rev.*, **141**, 2910–2945, doi:10.1175/MWR-D-12-00216.1.

Molteni, F., and Coauthors, 2011: The new ECMWF Seasonal Forecast System (System 4). ECMWF Tech. Memo. 656, 49 pp. [Available online at http://www.ecmwf.int/en/elibrary/11209-new-ecmwf-seasonal-forecast-system-system-4.]

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, doi:10.1175/MWR2906.1.

Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16–30, doi:10.1034/j.1600-0870.2003.201378.x.

Schefzik, R., T. L. Thorarinsdottir, and T. Gneiting, 2013: Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.*, **28**, 616–640, doi:10.1214/13-STS443.

Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **140**, 1086–1096, doi:10.1002/qj.2183.

Siegert, S., P. G. Sansom, and R. M. Williams, 2016: Parameter uncertainty in forecast recalibration. *Quart. J. Roy. Meteor. Soc.*, **142**, 1213–1221, doi:10.1002/qj.2716.

Sinha, A., G. Kathayat, H. Cheng, S. F. M. Breitenbach, M. Berkelhammer, M. Mudelsee, J. Biswas, and R. L. Edwards, 2015: Trends and oscillations in the Indian summer monsoon rainfall over the last two millennia. *Nat. Commun.*, **6**, 6309, doi:10.1038/ncomms7309.

Smith, D. M., A. A. Scaife, and B. P. Kirtman, 2012: What is the current state of scientific knowledge with regard to seasonal and decadal forecasting? *Environ. Res. Lett.*, **7**, 015602, doi:10.1088/1748-9326/7/1/015602.

——, R. Eade, and H. Pohlmann, 2013: A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction. *Climate Dyn.*, **41**, 3325–3338, doi:10.1007/s00382-013-1683-2.

Stephenson, D. B., C. A. S. Coelho, F. J. Doblas-Reyes, and M. Balmaseda, 2005: Forecast assimilation: A unified framework for the combination of multi-model weather and climate predictions. *Tellus*, **57A**, 253–264, doi:10.1111/j.1600-0870.2005.00110.x.

——, M. Collins, J. C. Rougier, and R. E. Chandler, 2012: Statistical problems in the probabilistic prediction of climate change. *Environmetrics*, **23**, 364–372, doi:10.1002/env.2153.

Stockdale, T. N., 1997: Coupled ocean–atmosphere forecasts in the presence of climate drift. *Mon. Wea. Rev.*, **125**, 809–818, doi:10.1175/1520-0493(1997)125<0809:COAFIT>2.0.CO;2.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, doi:10.1175/BAMS-D-11-00094.1.

Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. Roy. Soc. London*, **365A**, 2053–2075, doi:10.1098/rsta.2007.2076.

Trenberth, K. E., and Coauthors, 2007: Observations: Surface and atmospheric climate change. *Climate Change 2007: The Physical Science Basis*, S. Solomon et al., Eds., Cambridge University Press, 235–336.

Wang, M., and J. E. Overland, 2009: A sea ice free summer Arctic within 30 years? *Geophys. Res. Lett.*, **36**, L07502, doi:10.1029/2009GL037820.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2009: Seasonal ensemble forecasts: Are recalibrated single models better than multimodels? *Mon. Wea. Rev.*, **137**, 1460–1479, doi:10.1175/2008MWR2773.1.

Weijs, S. V., and N. van de Giesen, 2011: Accounting for observational uncertainty in forecast verification: An information-theoretical view on forecasts, observations, and truth. *Mon. Wea. Rev.*, **139**, 2156–2162, doi:10.1175/2011MWR3573.1.

Wilks, D. S., 2002: Smoothing forecast ensembles with fitted probability distributions. *Quart. J. Roy. Meteor. Soc.*, **128**, 2821–2836, doi:10.1256/qj.01.215.

——, 2006: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteor. Appl.*, **13**, 243–256, doi:10.1017/S1350482706002192.

——, 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. Elsevier, 704 pp.

Williams, R. M., C. A. T. Ferro, and F. Kwasniok, 2014: A comparison of ensemble post-processing methods for extreme events. *Quart. J. Roy. Meteor. Soc.*, **140**, 1112–1120, doi:10.1002/qj.2198.

WMO, 2007: Commission for basic systems: Extraordinary session. WMO Tech. Rep. 1017, 350 pp. [Available online at http://library.wmo.int/opac/index.php?lvl=notice_display&id=9680.]