Calibration and combination of dynamical seasonal forecasts to enhance the value of predicted probabilities for managing risk

John A. Dutton · Richard P. James · Jeremy D. Ross

Prescient Weather Ltd and The World Climate Service
200 Innovation Blvd Suite 229, State College, PA 16803
john.dutton@prescientweather.com

Authors' Copy

Final Publication Available in *Climate Dynamics*
in the Topical Collection on CFSv2

http://link.springer.com/article/10.1007/s00382-013-1764-2

DOI 10.1007/s00382-013-1764-2

or at http://link.springer.com

25 April 2013

Abstract Seasonal probability forecasts produced with numerical dynamics on supercomputers offer great potential value in managing risk and opportunity created by seasonal variability. The skill and reliability of contemporary forecast systems can be increased by calibration methods that use the historical performance of the forecast system to improve the ongoing real-time forecasts. Two calibration methods are applied to seasonal surface temperature forecasts of the U. S. National Weather Service, the European Centre for Medium Range Weather Forecasts, and to a World Climate Service multi-model ensemble created by combining those two forecasts with Bayesian methods. As expected, the multi-model is somewhat more skillful and more reliable than the original models taken alone. The potential value of the multimodel in decision making is illustrated with the profits achieved in simulated trading of a weather derivative. In addition to examining the seasonal models, the article demonstrates that calibrated probability forecasts of weekly average temperatures for leads of two to four weeks are also skillful and reliable. The conversion of ensemble forecasts into probability distributions of impact variables is illustrated with degree days derived from the temperature forecasts. Some issues related to loss of stationarity owing to long-term warming are considered. The main conclusion of the article is that properly calibrated probabilistic forecasts possess sufficient skill and reliability to contribute to effective decisions in government and business activities that are sensitive to intraseasonal and seasonal climate variability.

Keywords Seasonal forecast · Intraseasonal forecast · Probability forecast · Climate variability · Climate risk · Forecast calibration · Success ratio · Fraction correct · Forecast skill · Forecast reliability

This article is a contribution to the *Climate Dynamics* Topical Collection on the Climate Forecast System Version 2 (CFSv2), which is a coupled global climate model that was implemented by National Centers for Environmental Prediction (NCEP) in seasonal forecasting operations in March 2011. This Topical Collection is coordinated by Jin Huang, Arun Kumar, Jim Kinter and Annarita Mariotti.

1. Introduction

Regional climate variability on temporal scales of weeks, months, and seasons has significant impacts on a wide variety of public and private activities. Climate related risk and opportunity shape commercial and government planning and activity and may have profound and diverse effects on business revenues and costs.

Users of monthly or seasonal forecast services seek information that will help them anticipate the most likely climate variations, mitigate potential adverse effects, and take advantage of seasonal opportunity. The most effective weekly to seasonal forecasts are expressed in terms of probabilities, in part because that quantifies uncertainty and in part because probabilistic forecasts can be combined with business-specific financial metrics to optimize decisions and actions. Successful probability forecasts allow decision makers to estimate accurately the probability of success of their available options.

Recognizing the potential value of weather and climate forecasts, the United States and other nations make considerable investments in atmospheric and oceanic observations, in analyzing the data, and in attempting to predict future atmospheric events or statistics with numerical forecast systems. This process is remarkably successful in predicting the weather over the range of a few days, but progress has been less notable with numerical forecasts for a season or two ahead. The forecast range of a few weeks—between the initial value problem of weather prediction and the boundary value problem of seasonal prediction—has seemed to be even more challenging.

Here we consider three numerical surface temperature (t2m) ensemble forecast products:

- The second version of the Climate Forecast System (CFSv2) of the U.S. National Weather Service (NWS) (Saha et al. 2013),
- The fourth version of the Seasonal Forecast System (SFSv4) of the European Centre for Medium-Range Weather Forecasts (ECMWF) (Molteni et al. 2011), and
- A World Climate Service (WCS)[1] multi-model forecast created by a Bayesian combination of the CFSv2 and the ECMWFv4 forecasts.

For clarity, we will refer to these forecasts as CFSv2, ECMWFv4, and WCS MME, despite the evidently inconsistent naming convention.

The aim here is to demonstrate that the probabilistic versions of these forecasts have sufficient skill to be useful in planning and making decisions in both business and public activities. Thus we focus on measures of forecast quality that emphasize their utility in decision-making rather than some measures often used by other

[1] The World Climate Service is a commercial seasonal forecast service that provides a variety of climate information, analog construction tools, a monthly forecast and diagnostic document, and calibrated monthly forecasts of the NWS CFSv2, the ECMWF SFSv4, and the WCS multimodel ensemble at http://www.worldclimateservice.com. The WCS is a joint enterprise of Prescient Weather Ltd and MeteoGroup, an international independent weather and climate information provider in Europe, the U.S., and Asia.

authors. We also demonstrate that the WCS multi-model is generally superior to either of the two original models taken alone, as is anticipated by the present efforts in the U.S. and in Europe to create seasonal forecasts from independent prediction systems: The National Multi Model Ensemble (NMME) in the U.S. (Climate Prediction Center 2011a) and the Euro Seasonal-Interannual Prediction (EUROSIP) in Europe and the U.S. (ECMWF 2012), both motivated in part by the success of the DEMETER project (Palmer et al., 2004).

To date, analog and statistical approaches to seasonal prediction as discussed by van den Dool (2007) have proved generally superior to dynamical forecast methods; Livezey and Timofeyeva (2008) reached the same conclusion in a review of progress in U.S. seasonal forecasts. But the results here show that the probability forecasts constructed from the new versions of the computer models of weekly, monthly, and seasonal climate variations now have sufficient skill and reliability to be relevant and meaningful to climate-sensitive enterprises, either directly or as guidance considered in objective or subjective forecast methods.

2. Creating and Calibrating Probabilistic Seasonal Forecasts

Contemporary supercomputers can calculate tens of global seasonal forecasts simultaneously and thus initial conditions and model physics can be perturbed to generate an ensemble of forecasts whose evolving spread is expected to indicate the uncertainty of the forecast. Moreover, the ensemble members can be arranged to produce a probability distribution of predicted values of the variables at specific locations and times.

Forecast error generally increases as lead time increases from days to a season or two. Thus climate variability forecasts are assessed and improved by applying the forecast system to several decades of historical cases and comparing the results with the corresponding observations. The observed errors are summarized statistically to serve as corrections for future forecasts, with the local errors in mean values and various measures of spread being the focus of the calibration process.

The core of a seasonal probability forecast system is a dynamical Numerical Prediction System (NPS) that accepts initial and boundary conditions for the atmosphere and ocean and produces an ensemble of forecasts or analyses. The NPS is used as the engine to convert historical remote and in-situ observations about the atmosphere, ocean, and land and ice surfaces into a reanalysis dataset that represents a comprehensive climatology with atmospheric and ocean fields produced by the NPS from the observations. This reanalysis can then be used as the initial and boundary conditions for a large set of retrospective seasonal forecasts computed for, say, every month from 1980 to the present for leads of, say, one to six months. The reanalysis can serve as the verification observations for assessing the skill of the forecasts and developing calibration statistics. Of course, any comprehensive set of atmospheric and oceanic observations or independent reanalyses could be used as well for this purpose. Finally, the NPS is used to compute the actual seasonal forecasts from some time forward and a calibration process can be applied by the agency computing the forecast or by others, such as the WCS, that are creating value-added products from the numerical forecasts.

It is clear that the reanalysis, the retrospective forecasts, and the operational forecasts are equally important components of any process to produce calibrated seasonal forecasts. The critical assumption is that:

Past errors are a prolog to future errors and
can be used to improve future forecasts.

There is an evident assumption here that the evolving climate, the observations, and the NPS are all statistically stationary. Evidence of non-stationarity merits careful consideration of how it might be managed in the calibration process. We will encounter non-stationary conditions at various points in this article.

We use calibration schemes that simulate the actual operational forecast process for all the seasonal and intraseasonal forecasts discussed here. For the seasonal forecast, the historical period is 1982-1999 and forecasts are made for the ten years 2000 to 2009[2] with the calibration computed over an 18-year period prior to each forecast (Fig.1). The verification data for all three forecasts is obtained from the NCEP-DOE Reanalysis 2 (Climate Prediction Center, 2011b) and all model forecasts were interpolated to the 1.875 degree Gaussian grid of this reanalysis.

2.1 A variance adjustment calibration scheme

A first calibration scheme that provides a direct adjustment of bias and ensemble spread was based on ideas reported by Doblas-Reyes et al. (2005) and Johnson and Bowler (2009).

We denote a member of the forecast ensemble at some point on the spatial domain as

$$f_i(t) = F(t) + f_i'(t) \quad i = 1, 2, \cdots, N \tag{1}$$

in which $F(t)$ is the ensemble average and $f_i'(t)$ is a deviation from that average. The time variable will normally be discrete—a sequence of the same months for a number of years, for example. The verification data at the same spatial point for the same time period will be $v(t)$. We denote ensemble averages with $<()>$ and temporal averages in the domain $T_0 \le t < T$ with an overbar $\overline{(\,)}$.

Let the calibrated forecast be

$$\hat{f}_i(t) = \alpha F(t) + \beta f_i'(t) + F_0, \quad T_0 \le t \le T\,, i = 1, 2, \cdots, N \tag{2}$$

with spatially-dependent constants α, β, and F_0. In this formulation, the constant β scales the departure of the ensemble members from the ensemble mean, and the constant α adjusts the ensemble mean. If $\alpha = 1$, the constant F_0 simply represents a bias adjustment that ensures that the temporal averages of the ensemble mean and

[2] The CFSv2 became operational in March 2011. Historical forecasts for 2010 were not available when these computations were initiated.

verification are identical at each spatial point over the verification set. The calibrated ensemble mean is then a bias-adjusted ensemble mean.

One method of determining an optimal value for α is to minimize the mean-square error of the ensemble mean with respect to the verification. However, when this approach was applied to forecasts for 2000-2009 it was found that the calibrated ensemble mean showed degraded performance (larger mean error) than the bias-adjusted ensemble mean ($\alpha = 1$). It appears that the poorer performance is a result of evolving quasi-decadal trends in the forecast and verification datasets; consequently, optimizing α within the calibration history amounts to statistical overfitting, and the forecasts for subsequent years are less skillful than if no adjustment had been performed. Therefore the variance adjustment calibration proceeds here with $\alpha = 1$.

The appropriate scaling of the ensemble spread is determined by equating the ensemble variance with the variance of the observations within the calibration history

$$< \overline{(\hat{f}_i - \hat{F})^2} > = < \overline{(\beta f_i')^2} > = \beta^2 \sigma_{f'}^2 = \sigma_v^2 \tag{3}$$

which leads to

$$\beta^2 = \sigma_v^2 / \sigma_{f'}^2 \tag{4}$$

Once the calibrated ensemble members have been obtained, the probability of the predictand falling within any range may be estimated as the ratio of the number of ensemble members within the range to the total number of ensemble members. For example, the fraction of ensemble members for which the predictand is above the climatological normal provides a forecast of the probability of above-normal conditions.

More generally, we can produce a probability distribution for a predicted variable from the ensemble members at a point and a time. Let $\hat{f}_n, n = 1, 2, \ldots, N$ be the ensemble members sorted in increasing order. Then we can define the discrete probability distribution function

$$\text{Prob}\left[X \le \hat{f}_n \right] = P_X(\hat{f}_n) = n/(N+1) \tag{5}$$

2.2 The Gaussian Comb

A novel approach to ensemble forecast calibration was proposed by Raftery et al. (2005) in which the uncertainty surrounding each ensemble member is described by a probability density function centered on the predicted value. The individual density functions are then summed, with weights determined by a maximum likelihood algorithm that iterates on the forecast history, to obtain the overall predicted density. For temperature forecasts it is appropriate to use Gaussian density functions for the individual densities; the collection of densities then forms a set of "teeth" in what is known as a Gaussian comb.

The weights that are assigned to the individual member densities represent, in a Bayesian framework, the likelihood that each member is the best ensemble member. In the general case, the weights differ according to the skill of each member in the forecast history. However, if the ensemble members are known to be equivalent in terms of performance, the weights may be constrained to be equal.

A single parameter σ describes the variance of the individual member Gaussian density functions, and both the weights and σ are determined by the expectation maximization (EM) algorithm described below. After the iteration converges, σ can be further refined to ensure that the verifications fall within a specific probability range in the predicted historical ensemble, also described further below.

In applying the scheme to global seasonal forecasts, we allow the calibration parameters to vary in space because the computer models may have different performance characteristics in different parts of the world. Therefore we have computed the calibration parameters at each model grid point separately. This choice dictates that the ensemble member weights are constrained to be equal for all members of the same model, because the 18-year forecast history is not sufficiently long to compute different weights for each ensemble member. An alternative method would be to compute the parameters using all forecasts and observations within a certain radius of influence; this approach would avoid overfitting the weights, but would greatly increase the computational requirements.

The Gaussian comb was also applied to the WCS super-ensemble of CFSv2 and ECMWFv4 forecasts. The member weights were constrained to be equal within each model's ensemble, but different weights were permitted for the two models, thereby allowing the locally superior model to attain greater influence in the calibrated forecasts. Thus the method determines the standard deviation σ and the weights α and $1-\alpha$ for the two models. We will refer to this as the Bayesian calibration of the WCS multi-model and note that the variance adjustment algorithm is not used in this process.

The combined probability functions describing the Gaussian comb forecast for a predictand y are

$$p(y) = \sum_{n=1}^{N} w_n \frac{1}{\sqrt{2\pi}\,\sigma} \exp[-\frac{1}{2}(\frac{y-z_n}{\sigma})^2], \qquad \sum_{n=1}^{N} w_n = 1$$

$$P(y) = \int_{-\infty}^{y} p(\eta)\,d\eta = \sum_{n=1}^{N} w_n \frac{1}{2} \mathrm{erfc}[-\frac{y-z_n}{\sqrt{2}\,\sigma}] \tag{6}$$

in which the z_n are the bias-adjusted forecasts from the N-member ensemble, the w_n are the weights of the ensemble members, and σ is the variance of the individual density functions in the Gaussian comb.

The expectation maximization algorithm (Dempster et al. 1977) proceeds by iterating over the available history according to the following scheme until a satisfactory degree of convergence is achieved:

$$
\begin{aligned}
\eta_{n,t} &= w_n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{y_t - z_{n,t}}{\sigma})^2} \\
\eta_{n,t} &\rightarrow \eta_{n,t} / \sum_{j=1}^{N} \eta_{j,t}
\end{aligned}
\tag{7}
$$

for $n = 1, 2, \ldots, N$ and then

$$
\begin{aligned}
w_n &= \frac{1}{T} \sum_{t=1}^{T} \eta_{n,t} \\
\sigma^2 &= \frac{1}{T} \sum_{n=1}^{N} \sum_{t=1}^{T} \eta_{n,t} (y_t - z_{n,t})^2
\end{aligned}
\tag{8}
$$

The probability functions in (6) provide numerical values once (8) converges to values for the weights and the standard deviation.

2.3 Three Measures of Forecast Quality

In this section, we apply three measures of the quality of the seasonal forecasts to evaluate and compare several sets of calibrated seasonal forecasts[3].

2.3.1 Mean Absolute Errors

The first is the mean absolute error (MAE) between the forecasts and the verification data averaged over the forecast domain and over some set of forecasts. Throughout this article, we consider the October forecasts for the season December, January, and February (DJF) and the April forecast for the season June, July, and August (JJA)—a two-month lead forecast for winter and summer in both hemispheres. The MAE errors are computed with the ensemble means and do not take advantage of the probabilistic information in the forecast ensemble.

The MAE statistics of the ensemble mean forecasts in Table 1 illustrate the very marginal skill of ensemble-mean seasonal climate forecasts. On a global basis both the CFSv2 and ECMWFv4 models show a marginal degree of skill compared to climatology in Northern Hemisphere winter, but are not skillful in summer. Root mean square errors (RMSE) give an impression of slightly greater skill relative to climatology but are not shown here. The general lack of meaningful skill relative to climatology is well known to climate forecasters but does not imply that the model forecasts are without value; rather, the distribution of forecast members within the model ensemble must be utilized to create meaningful guidance in probabilistic form.

[3] Our results may appear to differ from other studies of these two models(e.g., Kim, Webster, and Curry, 2012), but it is essential to observe that measures of quality will depend on model calibration methods and the verification datasets used. Kim et al. focus on model bias; we focus on the probabilistic forecasts.

2.3.2 Prediction Interval Coverage

A second simple but useful measure of the quality of a forecast calibration scheme is the coverage of a prediction interval by the calibrated forecast ensemble. The prediction interval for a forecast is the range (centered on the expected value) determined from the forecast probability distribution within which the predictand (*e.g.* temperature) is expected to fall with a certain probability; for example, a 66.7 percent prediction interval is expected to contain the verifying observation with 66.7 percent frequency and in the case of a well-calibrated forecast will indeed contain the predictand on approximately 66.7 percent of occasions. Given a calibrated forecast, then, it is possible to count the fraction ("coverage") of observations that fall within a prediction interval and thereby assess the goodness of the calibration. Furthermore, it is possible to perform a further calibration to adjust the spread of the ensemble members to cover the desired prediction interval more adequately.

The widths and observed coverage of the 66.7 percent prediction intervals from the unadjusted model ensembles, the bias-adjusted ensembles, and the calibrated ensembles are shown in Table 2.

The prediction intervals from the original model ensemble contain notably too few observations in all cases, but this problem is partly caused by the model bias. After applying the bias adjustment, the prediction interval remains too small on a global basis, indicating that the ensemble spread is too small on average. The calibrated prediction interval, in contrast, is fairly close to capturing the correct fraction of observations, indicating that the calibration schemes are performing reasonably well.

2.3.3 Reliability Diagrams

A comparison of predicted probabilities and observed frequencies provides a third measure of forecast quality. We determine boundaries for below, near, and above normal categories with the verification data in the calibration set; the categories are often but not necessarily terciles. Then using the ensemble probabilities, we obtain predicted probabilities for the observation occurring in each category for each forecast point and time. We divide the predicted probability axis into a set of bins and record the number of predictions that fall in each bin along with the number of cases in which the forecast was correct. The total number of forecasts falling in each bin is known as the sharpness $s(p)$ of the forecast; the number of correct forecasts in each corresponding bin is the verification $v(p)$. For each of the three possibilities of below, near and above normal, we define the reliability of the forecast to be

$$r(p) = v(p)/s(p) \tag{9}$$

and plot $r(p)$ against p as the observed frequency of the events we are attempting to predict. For a perfect probability forecast, the events predicted to occur with probability p will occur with a frequency p and so we will have $r(p) = p$ and a plot of $r(p)$ will lie along the diagonal. Usually probability forecasts will be too confident at high probabilities and not sufficiently confident at low probabilities and thus the reliability curve $r(p)$ will be flatter than the diagonal. Various aspects of the reliability diagram are discussed by Wilks (2006).

Reliability diagrams for the October forecasts for winter (DJF) for 2000-2009 are shown in Fig. 2 for the variance adjustment and in Fig. 3 for the Bayesian calibration. While it is evident subjectively from the diagrams that the WCS MME forecasts are more reliable than the individual forecasts, it is worth attempting a quantitative measure of reliability to provide an objective comparison. Perfect reliability curves have a slope of 1 while the curve for forecasts using the climatological probability would be flat near a value expected to be near one-third for ternary forecasts and to have a zero slope. We define a reliability index as the best-fit linear slope s of the reliability curve with a penalty for empty bins as $RI = s\,f$ in which f is the fraction of bins that have estimates[4]. Thus $RI = 1$ for perfect reliability and $RI = 0$ for climatology. The values of this index for the reliability curves in Figs. 2 and 3 are provided in Table 3 and verify the visual impression that the WCS MME forecasts are the most reliable of the three forecasts.

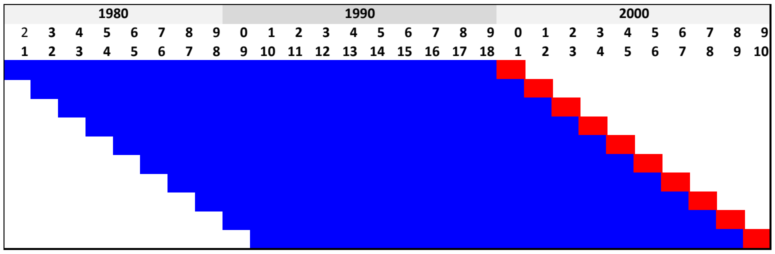


Fig. 1 Scheme for calibrating simulated operational forecasts, with the historical calibration period in blue and the forecast month in red.

[4] Some of the empty bins actually contained a small number of correct high probability forecasts that were eliminated by the criterion that bins must have 20 total forecasts or more, suggesting that longer historical periods might be advantageous.

Table 1 Mean absolute error of lead 2-4 month CFSv2 and ECMWFv4 seasonal 2 m temperature forecasts for DJF and JJA, 2000-2009 for the original and bias-adjusted (bias-adj) ensembles. The top row gives the mean-absolute difference between individual reanalysis values and reanalysis climatology (°C). All other rows give the ratio of the MAE for the forecasts to the MAE for climatology. Areas are Global (GL), North America (NA), Europe (EU) and Tropical Pacific (TP) delineated by rectangles covering the areas named, thus including some ocean points for NA and EU.

	October for Winter (DJF) 2000-2009				April for Summer (JJA) 2000-2009			
Model	GL	NA	EU	TP	GL	NA	EU	TP
Climatology	1.01	1.16	1.20	0.54	0.74	0.70	0.67	0.34
CFSv2	1.64	1.33	1.59	1.28	2.23	2.36	3.09	1.32
CFSv2 bias-adj	0.96	0.95	0.97	1.04	1.03	1.04	1.01	1.12
ECMWFv4	1.86	1.87	1.53	3.17	2.27	2.04	1.61	5.35
ECMWFv4 bias-adj	0.96	0.96	0.99	0.57	0.99	1.03	1.0	0.88
MME	1.61	1.46	1.48	2.00	2.07	1.93	2.06	3.06
MME bias-adj	0.92	0.91	0.97	0.69	0.97	1.00	0.99	0.88

Table 2 Fraction of verifying observations falling within the 66.7 percent prediction interval of leads 2-4 month for CFSv2 and ECMWFv4 seasonal 2 m temperature forecasts for DJF and JJA, 2000-2009. The top table gives width of the interval in the climatology and then the ratio of the width for the various models to the climatology width. The bottom table gives the ratio of the coverage for the 67 percent interval to 67 per cent. Abbreviations: Multi-Model Ensemble MME, Global GL, North America NA, Europe EU, Tropical Pacific TP, Bias Adjusted Bias-Adj , Variance Adjustment Var Adj, and Bayesian Bayes.

	October for Winter (DJF) 2000-2009				April for Summer (JJA) 2000-2009			
Width of 67% Prediction Interval								
Model	GL	NA	EU	TP	GL	NA	EU	TP
Climatology (°C)	2.3	2.84	3.05	1.4	1.88	1.8	1.73	1.03
CFSv2 Bias-Adj	0.77	0.99	0.90	0.31	0.73	0.71	0.86	0.50
CFSv2 Var Adj	0.93	0.93	0.90	0.92	0.97	0.95	0.98	1.08
CFSv2 Bayes	0.94	1.00	0.95	0.65	0.95	0.90	0.97	0.78
ECMWFv4 Bias-Adj	0.79	1.04	0.93	0.37	0.81	0.91	0.97	0.45
ECMWFv4 Var Adj	1.00	1.04	0.97	1.06	1.02	1.06	0.99	1.01
ECMWFv4 Bayes	0.94	1.04	0.99	0.51	0.95	0.98	1.02	0.72
WCS MME Bayes	0.92	1.00	0.94	0.56	0.93	0.93	0.96	0.74
Verification Coverage of 67% Prediction Interval								
Model	GL	NA	EU	TP	GL	NA	EU	TP
Climatology	0.96	0.99	0.97	1.01	1.00	1.01	1.04	1.03
CFSv2 Bias-Adj	0.82	1.00	0.96	0.31	0.79	0.82	0.93	0.63
CFSv2 Var Adj	0.94	0.94	0.94	0.94	0.97	0.99	1.01	1.12
CFSv2 Bayes	0.94	1.01	1.00	0.67	0.96	0.96	1.01	0.93
ECMWFv4 Bias-Adj	0.84	1.04	0.91	0.67	0.84	0.96	1.01	0.64
ECMWFv4 Var Adj	1.00	1.01	0.93	1.36	1.03	1.04	1.01	1.18
ECMWFv4 Bayes	0.96	1.06	0.97	0.93	0.97	1.00	1.04	0.97
WCS MME Bayes	0.96	1.07	0.97	0.85	0.99	1.00	1.04	0.99

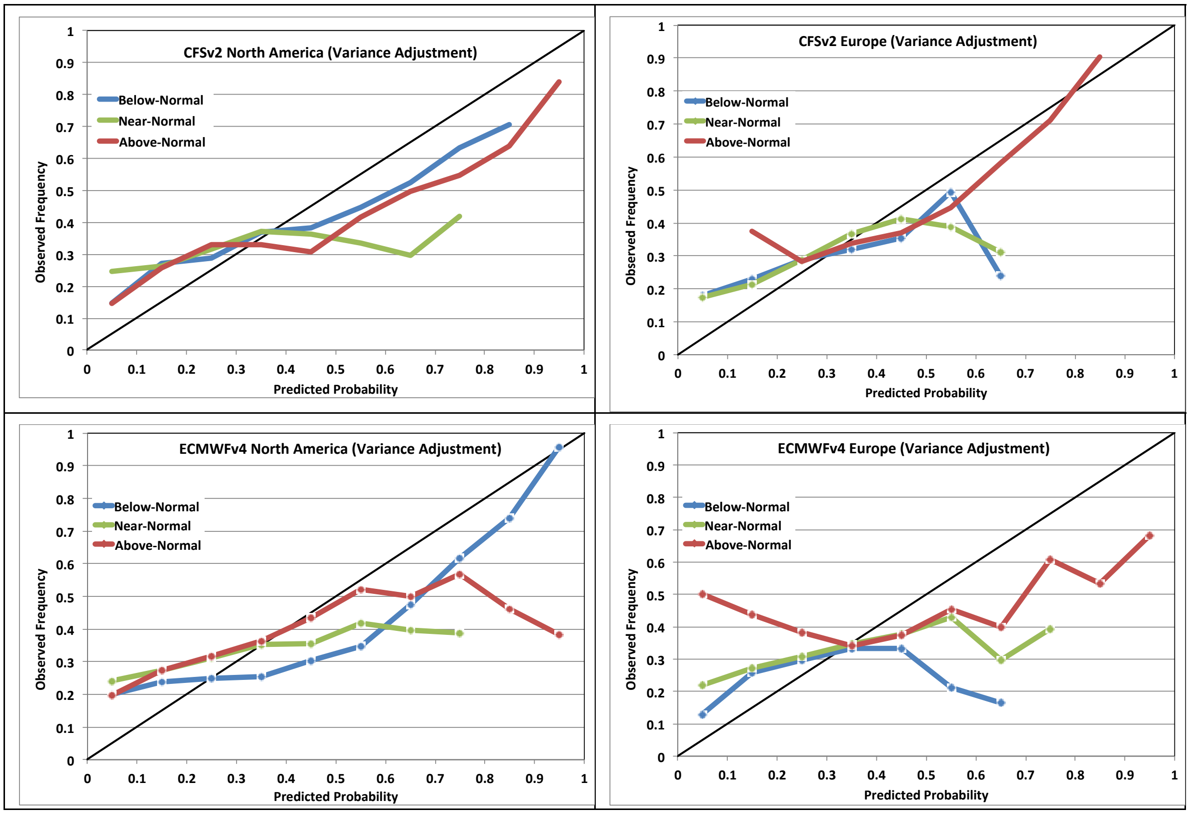
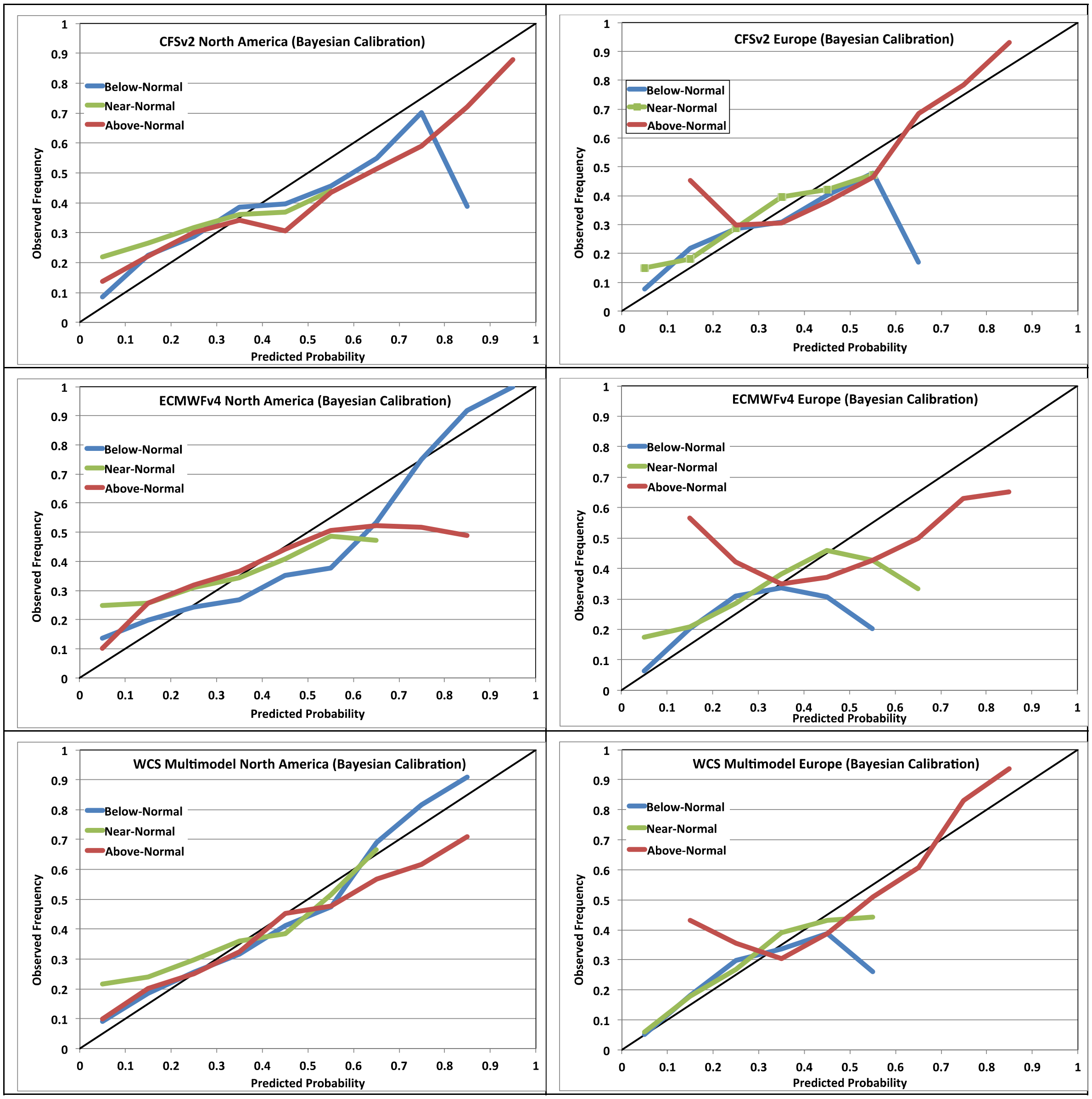


Fig. 2 Reliability diagrams for the CFSv2 and ECMWFv4 lead 2-4 forecasts for winter calibrated with the variance adjustment algorithm. Sharpness bins with less than 20 forecasts are considered statistically insignificant and are neglected.



CFSv2 North America (Bayesian Calibration)

CFSv2 Europe (Bayesian Calibration)

ECMWFv4 North America (Bayesian Calibration)

ECMWFv4 Europe (Bayesian Calibration)

WCS Multimodel North America (Bayesian Calibration)

WCS Multimodel Europe (Bayesian Calibration)

Fig. 3 Reliability diagrams for the CFSv2, ECMWFv4, and WCS MME lead 2-4 forecasts for winter calibrated with the Bayesian algorithm. Sharpness bins with less than 20 forecasts are considered statistically insignificant and are neglected.

Table 3 Reliability indexes (in percent) for the lead 2-4 forecasts for winter (DJF) and summer (JJA) for 2000-2009 confirming that the WCS MME forecasts are generally more reliable than the CFSv2 and ECMWFv4 forecasts taken individually. Abbreviations: Below Normal B, Near Normal N, Above Normal A, Variance adjustment calibration Var Adj, Bayesian model calibration Bayes.

Model	Method	Average of B N A	Below Normal	Near Normal	Above Normal	Average of B N A	Below Normal	Near Normal	Above Normal
		North America Winter (DJF) 2000-2009				Europe Winter (DJF) 2000-2009			
CFSv2	Var Adjust	***45***	58	14	64	***35***	19	22	63
CFSv2	Bayes	***49***	48	25	74	***44***	23	42	67
ECMWFv4	Var Adjust	***42***	78	19	29	***13***	1	17	22
ECMWFv4	Bayes	***57***	98	31	43	***22***	18	27	21
WCS MME	Bayes	***70***	93	50	67	***49***	34	48	66
		North America Summer (JJA) 2000-2009				Europe Summer JJA 2000-2009			
CFSv2	Var Adj	***41***	43	28	51	***26***	18	16	44
CFSv2	Bayes	***52***	56	42	57	***31***	33	28	31
ECMWFv4	Var Adj	***41***	51	27	46	***46***	1	39	34
ECMWFv4	Bayes	***53***	74	39	47	***29***	13	42	33
WCS MME	Bayes	***63***	77	49	64	***33***	20	38	42

3. Skill of Probabilistic Forecasts for Decision-Making

Users of seasonal forecasts can mitigate risk or take advantage of opportunity implied by the predicted probability of events to come. For example, electric utilities can smooth financial performance by hedging with weather derivatives against expectations of warm winters or cool summers.

A contingency table comparing forecasts to observations allows us to explore the potential application of seasonal forecasts in this manner. The various cases are represented in Table 4. Often all quantities in the contingency table are converted to frequencies by dividing them by the total number *N* of forecasts.

In the table, *a* is the number of forecasts of above normal conditions that verified as above normal, *b* the number that verified as normal, and *c* those that verified as below normal. The f_i are the numbers of forecasts in the terciles, the n_i the numbers of observations.

We define success ratios of the form $S_a = a / n_a$ as the fraction of events predicted correctly, and we define the fractions correct $F_a = a / f_a$ as the fraction of forecasts that were correct. With similar definitions for the other two terciles, we define the quantities in Table 5. The fractions correct and the success ratios will be equal if the contingency table is symmetric about the main diagonal. Both will have values of one-third for random forecasts for ternary events.

The fractions of correct forecasts and the success ratios for the three models are shown in Tables 6 and 7, with the best model for each geographical area and time period indicated by bold italic type. The relative performance of the three models is summarized in Table 8 which shows that the WCS MME is slightly better for fractions correct.

Table 4 Contingency table for examining the skill of probabilistic forecasts for ternary events.

		Observations			
		Above Normal	Near Normal	Below Normal	Number of Forecasts
Forecasts	Above Normal	*a*	*b*	*c*	f_a
	Near Normal	*d*	*e*	*f*	f_n
	Below Normal	*g*	*h*	*i*	f_b
	Number of Observations	n_a	n_n	n_b	*N*

Table 5 Definitions of summary measures of forecast skill for ternary events. All quantities in Table 4 have been scaled here by *N* so that $n_a + n_n + n_b = f_a + f_n + f_b = 1$

Fraction of Correct Forecasts	$F = f_a F_a + f_n F_n + f_b F_b$
Fraction of Events Predicted Correctly	$S = n_a S_a + n_n S_n + n_b S_b = F$
Perfect Forecasts	$F_p = S_p = 1$
Random Forecasts	$F_r = S_r = 1/3$
Improvement Ratios	$(F - F_r)/F_r, \quad (S - S_r)/S_r$

Table 6 Fractions of correct forecasts (in percent) for Bayesian calibration. The best forecast for each geographic region, each tercile, and for the diagonal elements taken together (denoted as Total) is identified with bold type. Random ternary forecasts would have fractions correct of 33 percent.

	October Forecasts for Winter 2000-2009				April Forecasts for Summer 2000-2009			
Model	Below Normal	Near Normal	Above Normal	Total	Below Normal	Near Normal	Above Normal	Total
				NORTH AMERICA				
CFSv2	43	37	44	43	39	44	43	42
ECMWFv4	40	***45***	46	44	41	**47**	43	43
WCS MME	***45***	42	***48***	***46***	***42***	45	43	***44***
				EUROPE				
CFSv2	38	***44***	***44***	***43***	***33***	37	47	44
ECMWFv4	31	41	41	39	21	***47***	48	45
WCS MME	39	41	43	42	32	43	48	***46***
				TROPICAL PACIFIC				
CFSv2	81	33	48	48	***67***	42	51	53
ECMWFv4	79	***48***	***69***	***64***	59	44	***59***	55
WCS MME	***85***	44	61	59	66	***47***	56	***56***
				GLOBAL				
CFSv2	47	40	52	49	30	42	51	45
ECMWFv4	41	***42***	53	47	30	42	52	44
WCS MME	***48***	41	***54***	***50***	***32***	***43***	***52***	***46***

Table 7 Success ratio for forecasts (in percent for Bayesian calibration), as in Table 6. Random ternary forecasts would have success ratios of 33 percent.

	October Forecast for Winter				April Forecasts for Summer			
Model	Below Normal	Near Normal	Above Normal	Total	Below Normal	Near Normal	Above Normal	Total
				NORTH AMERICA				
CFSv2	33	19	***71***	43	27	27	68	42
ECMWFv4	***50***	***26***	54	44	***34***	22	70	43
WCS MME	42	24	68	***46***	27	***28***	71	***44***
				EUROPE				
CFSv2	***21***	27	73	***43***	12	19	77	44
ECMWFv4	19	25	65	39	***13***	23	75	45
WCS MME	14	***28***	73	42	8	23	***81***	***46***
				TROPICAL PACIFIC				
CFSv2	29	23	87	48	44	16	***94***	53
ECMWFv4	***47***	***51***	90	***64***	***57***	***33***	74	55
WCS MME	36	43	***92***	59	49	26	90	***56***
				GLOBAL				
CFSv2	39	25	***72***	49	28	27	***66***	45
ECMWFv4	***40***	31	64	47	***36***	32	57	44
WCS MME	38	31	70	***50***	29	***33***	63	***46***

Table 8 Average performance ratios for the above normal and below normal terciles, Bayesian calibration. The best ratios for each region are denoted by bold italic type.

	October for Winter 2000-2009			April for Summer 2000-2009		
Model	NA EU	TP	G	NA EU	TP	G
			FRACTIONS CORRECT			
CFS	42	65	50	41	59	41
ECMWF	40	***74***	47	38	59	41
WCS MME	***44***	73	***51***	41	***61***	***42***
			SUCCESS RATIOS			
CFS	***50***	58	***56***	46	69	47
ECMWF	47	***69***	52	***48***	66	47
WCS MME	49	64	54	47	***70***	46

4. Profits on Options Illustrate the Value of Seasonal Forecasts

A simple scheme involving options on weather or climate conditions will illustrate how the skill of the forecast can provide favorable financial performance[5]. Consider a hypothetical option on seasonal temperature averaging below, near, or above normal. Suppose an option for each case can be purchased for a preseason price of P and that the option will pay $3P$ to holders of options for the case that verified. There is an implicit assumption here that the market is large enough and the option purchases are sufficiently symmetric to support the contracts; for simplicity we ignore transaction costs.

For F the fraction of correct forecasts, the rate of return is

$$\begin{aligned} R &= (3FP - P)/P \\ &= 3F - 1 \\ &= (F - F_r)/F_r \end{aligned} \tag{10}$$

with the last line the improvement ratio with respect to random forecasts defined in Table 5.

Table 9 illustrates the returns on the hypothetical options that would have been achieved with the WCS multimodel forecasts, assuming that options were purchased for the tercile with the largest probability. An alternate strategy is to take a position each time the predicted probability for a tercile exceeds some threshold value.

To explore this concept, we define as cumulative quantities for all probabilities greater than p the number of forecasts, $S(p)$, the number of correct forecasts, $V(p)$, and the net financial gain from the trades, $G(p)$, with

$$S(p) = \int_p^1 s(y)dy, \quad V(p) = \int_p^1 v(y)dy, \quad G(p) = 3V(p) - S(y) \tag{11}$$

We use the definitions associated with (9) and then convert (10) from rate of return to gain. To be specific, we will assume that options worth \$1 M are purchased when the probability threshold is equaled or exceeded. Using the data associated with the reliability diagram for the WCS MME for North America for winter in Fig. 3, we compute (11) as sums over the probability bins and display the results in Fig. 4, which shows that maximum total net gains per average station would have been achieved by purchasing options whenever the predicted probabilities exceeded 40 per cent. For the below normal case, the investment in 2.3 options per station over the 10-year forecast period would have produced a net gain of \$1M and the above normal 5 options per station would have produced a net gain of \$2.4M with corresponding net returns of 43 and 48 percent. The difference in the two cases is created by the larger fraction of above normal forecasts accumulating at larger probabilities—a bias that we will examine in Section 7.

[5] Taylor and Buizza (2006) applied a similar approach, using ten-day ensemble forecasts to determine the pay-off probabilities of weather derivatives.

Table 9 Rate of return from trading the hypothetical weather derivative with the aid of the WCS MME for all forecasts in 2000-2009 for DJF and JJA.

	Global	North America	Europe
Fraction Correct	0.48	0.44	0.44
Return rate (per cent)	**44**	**32**	**32**

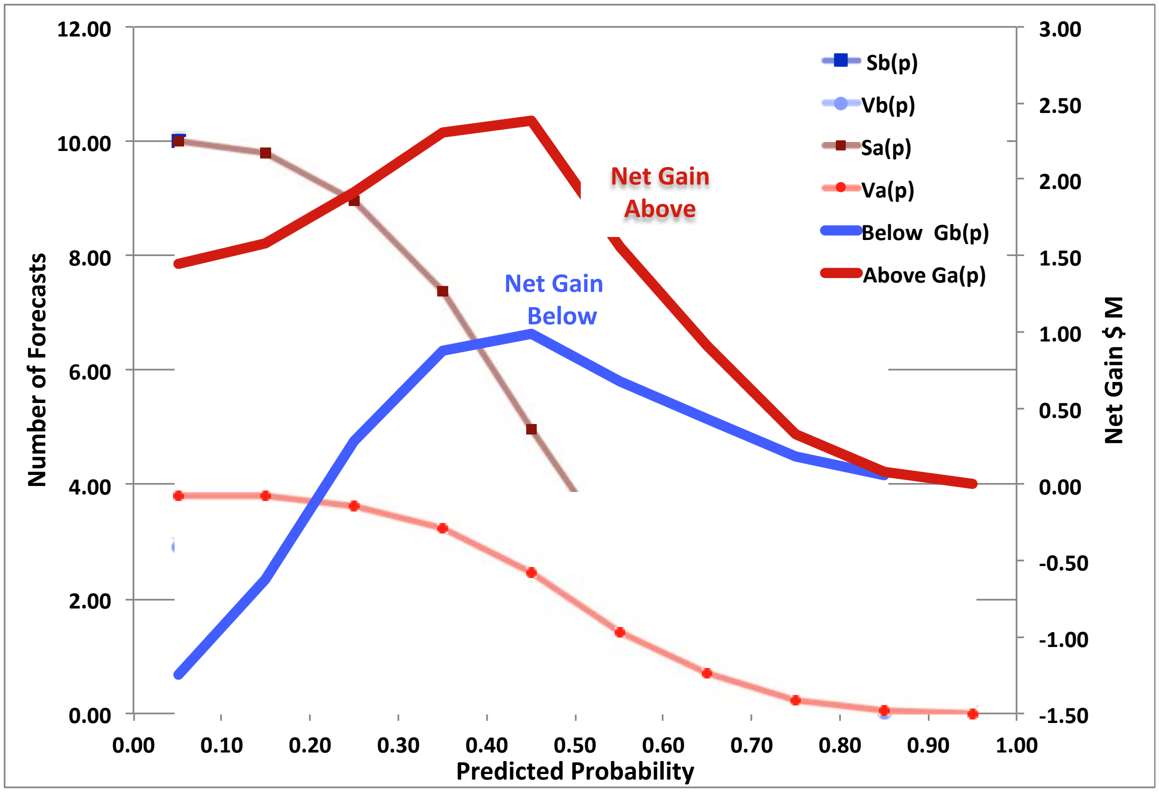


Fig. 4 Illustration of average gains per station obtained by investing in the hypothetical options when predicted probabilities of above or below normal conditions in the North American winter exceeded various predicted probabilities. The quantities accumulating for all probabilities exceeding *p* are defined in (11).

5. Intraseasonal Forecasts

Considerable commercial interest is focused on forecasts with lead times of two to six weeks, which hold the potential to dramatically influence tactical planning and risk management for weather-sensitive enterprises. Both the NWS and ECMWF run versions of their forecast systems designed to supply guidance on the intraseasonal timescale; for example, the ECMWF SFS generates a 30-day forecast twice per week along with a retrospective set of forecasts initialized on the same day of the year in each of the past 18 years.

The performance of the CFS and ECMWF intraseasonal forecasts was examined by applying the Gaussian comb calibration scheme to a complete set of retrospective forecasts from both models, and to the multi-model forecast (MME) obtained by combining the two ensembles. The calibration process was similar to the seasonal Gaussian comb calibration, except that the CFS Reanalysis (Saha et al. 2010) was used as the verification dataset, and the calibration and verification steps were performed over the periods of available daily data, 2000-2006 and 2007-2010 respectively. Forecasts of weekly-average 2-m temperature were tested for lead times of one to four weeks. Calibration parameters were computed separately for each month using all forecasts initialized within that month.

The MME forecasts showed good calibration characteristics, as illustrated by the reliability diagrams and indexes in Figure 5 and Table 10 respectively. The skill statistics for 2007-2010 also reveal that the calibrated intraseasonal forecasts possess significant skill and value even at the four-week lead time, which is often thought to pose a very challenging prediction problem. A summary of forecasts correct is shown in Table 11. As with the seasonal forecasts, the utility of the forecasts is derived mainly from the probabilistic information contained in the ensemble distribution, rather than from the relatively unskillful ensemble mean.

An example of the probabilistic forecast output produced by an operational version of the calibration scheme is shown in Fig. 6, which depicts an increasingly confident CFSv2 forecast for unusual warmth in early December 2012 across much of the continental US.

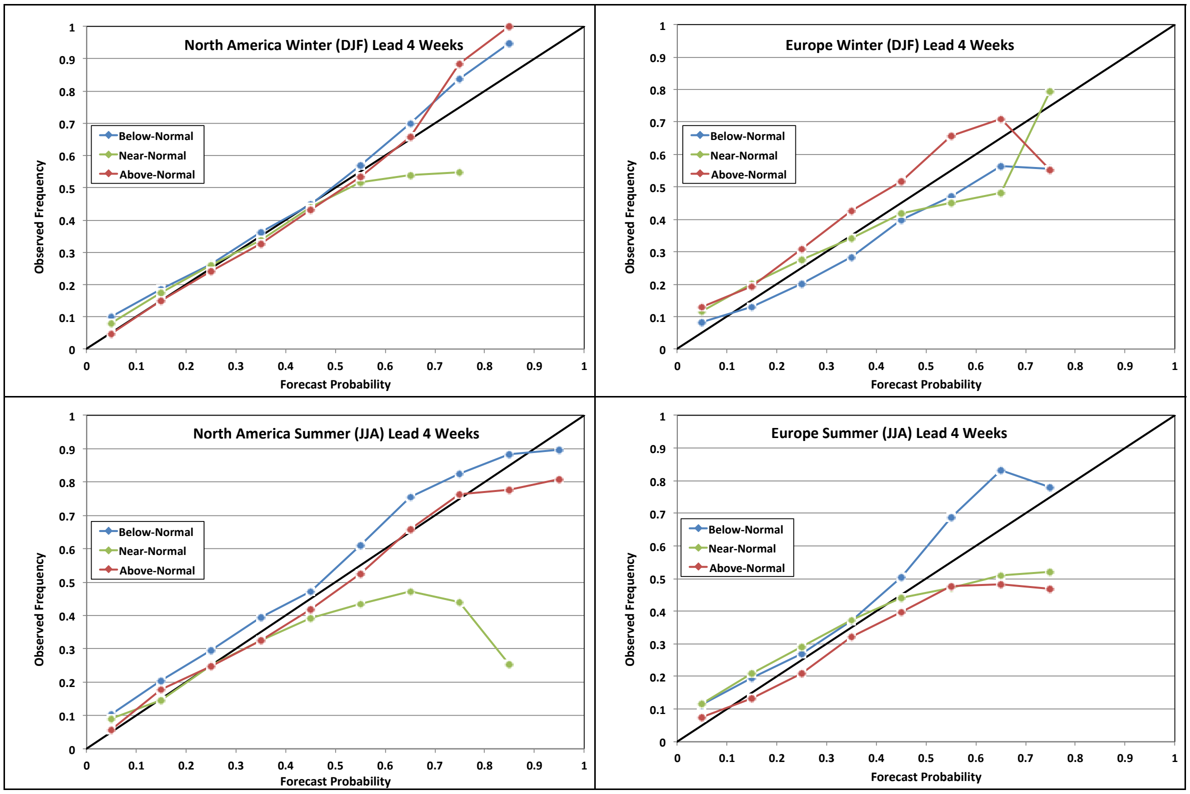


Fig. 5 Reliability diagrams for the WCS MME forecasts for leads of four weeks for North America and Europe for the winter (DJF) and summer (JJA) for 2007-2010. Sharpness bins with less than 20 forecasts are considered statistically insignificant and are dropped.

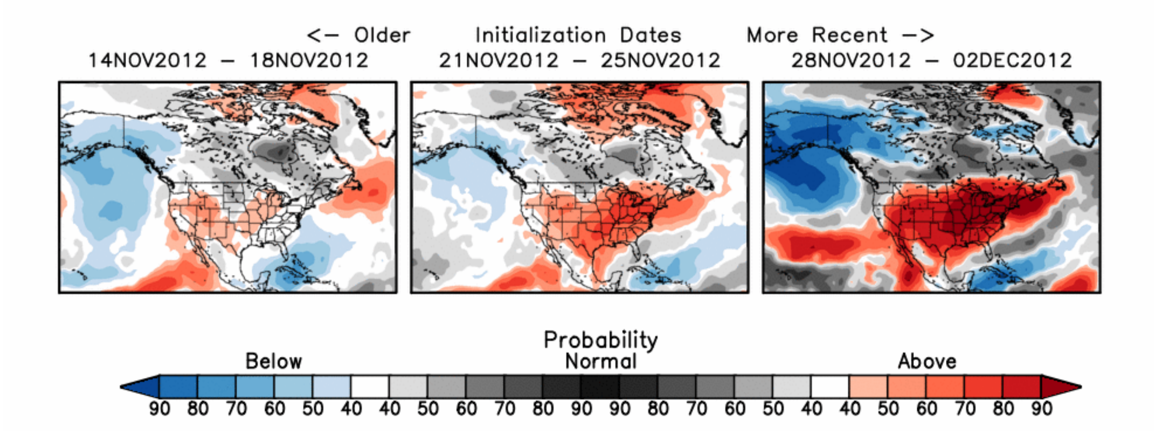


Fig. 6 CFSv2 probability forecast of 2-m temperature terciles over North America, valid for the week of 3-9 December 2012. The left panel shows the forecast from CFSv2 runs initialized in the period 14-18 November (~3 week lead), the middle panel forecasts initialized on 21-25 November (~2 week lead), and the right panel forecasts initialized on 28 November – 2 December (~1 week lead).

Table 10 Reliability indexes (in percent) for the WCS MME week 4 forecasts for winter (DJF) and summer (JJA) for 2007-2010. Abbreviations: Below Normal B, Near Normal N, Above Normal A. All forecasts were calibrated with the Bayesian algorithm.

Model	Average of B N A	Below Normal	Near Normal	Above Normal	Average of B N A	Below Normal	Near Normal	Above Normal
	North America Winter Week 4 2007-2010				Europe Winter Week 4 2007-2010			
CFSv2	**83**	87	97	64	**76**	68	66	94
ECMWFv4	**52**	57	55	45	**33**	47	30	22
WCS MME	**86**	97	57	106	**47**	55	29	56
	North America Summer Week 4 2007-2010				Europe Summer Week 4 2007-2010			
CFSv2	**68**	80	46	79	**64**	76	57	60
ECMWFv4	**64**	82	30	78	**52**	66	44	46
WCS MME	**72**	95	31	89	**62**	88	47	51

Table 11 Fractions correct for the WCS MME weekly forecasts 2007-2010 for below, near, and above normal categories averaged together.

	North America		Europe	
	Winter (DJF)	Summer (JJA)	Winter (DJF)	Summer (JJA)
Week 1	78	73	79	74
Week 2	61	56	57	55
Week 3	52	48	47	45
Week 4	48	45	44	42

6. Conversion to Impact Variables

Many decisions about responses to seasonal climate variability focus on quantities derived from observed or predicted meteorological variables. Examples include degree days in the energy and agriculture industries, wind power in energy, and moisture and evaporation indexes in hydro-energy, agriculture, and prediction of wildfire risk.

In such cases we may be interested in a variable y derived from a traditional atmospheric variable x by a possibly non-linear function $y(x)$ and we would like to find the probability distribution $P_y(Y)$ for the probability that $y \le Y$. For those cases in which $y(x)$ is continuous and monotonic, we would have an inverse function $x(y) = y^{-1}(y)$ and be able to compute $P_y(Y) = P_x(x(Y))$ from the predicted probability $P_x(X)$. Even in such cases, it is often more efficient to generate the statistics for y numerically. Then we simply convert the ensemble $\mathbf{x} = \{x_1, x_2, \cdots, x_N\}$ of forecasts of x into an ensemble $\mathbf{y} = \{y(x_1), y(x_2), \cdots, y(x_N)\}$ of predicted variables and proceed with the statistical analysis in analogy with (5).

We offer an example using degree days, which are widely used in the energy and agriculture industries. In 1999 the Chicago Mercantile Exchange introduced exchange-traded degree day futures that now encompass 47 cities worldwide, and thus it is becoming increasingly advantageous to serve energy and agriculture traders or risk managers by transforming seasonal model forecasts of temperature into variables such as heating (HDD), cooling (CDD), or growing (GDD) degree days. Because degree days are computed from the average of daily maximum and minimum temperatures, it is preferable to acquire daily temperature maxima and minima from high temporal resolution model forecasts. This requires significantly more data storage and computational resources, and as a result, quantitative estimates of degree days in commercial seasonal forecasts are rare.

To compute calibrated probability forecasts of degree days requires analysis of both the six-hourly forecasts and the observations four times a day in the set of historical forecasts in order to determine the bias correction. Then daily temperature maxima and minima in the unbiased forecast temperatures can be converted to degree days and the probability distributions determined. Figs. 7 and 8 show how the predicted probability distributions of cooling degree days and degree-day anomalies evolve over a summer month.

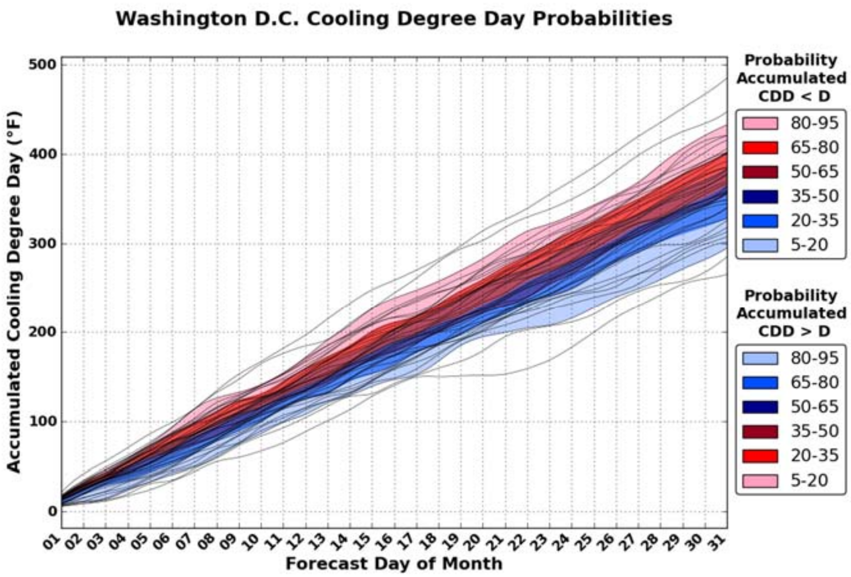


Fig. 7 Predicted probability distributions for cooling degree days for Washington D. C. for July 2012 created with CFSv2 forecasts in June.

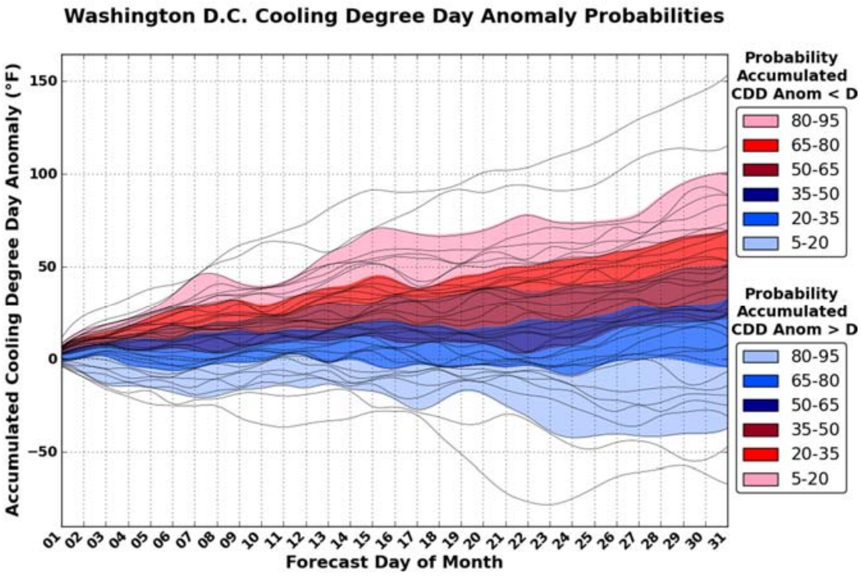


Fig. 8 Predicted probability distributions for cooling degree-day anomalies for Washington, D.C. for July 2012 created with the CFSv2 system in June.

7. Apparent Troubles with Long-Term Trends

The seasonal forecasts considered here exhibit an unexpectedly strong tendency to favor warmer than normal at the expense of cooler than normal. It is well known that long-term climate trends and variations owing to the El Niño-Southern Oscillation (ENSO) provide an important source of predictability for seasonal forecasts (e.g., van den Dool 2007, Livezey and Timofeyeva 2008), but the bias in the numerical forecasts considered here seems to be excessive and undesirable.

To demonstrate the problem, we recall that the tercile boundaries are determined for each grid point from the 18 seasonal averages in the historical verification set. These boundaries determine the distribution $\{f_a, f_n, f_b\}$ of the forecasts and the verification data $\{n_a, n_n, n_b\}$ using the conventions of the contingency table (Table 4). The results are summarized in Table 12 which shows that the fraction of above normal temperatures in the reanalysis and forecasts increases during the decade of forecasts while the number of below normal decreases. The skewing of the distributions is dramatically more pronounced for the forecasts. This skewing in present in both the CFSv2 and ECMWFv4 forecasts, but is not shown here for the individual models.

These statistics, it turns out, depend on relatively small variations of temperature. The width of the near normal temperature tercile for the zonally averaged surface temperature is less than 0.5 C in summer and less than 0.75 C in winter over most of the range between 60S and 30 N, as shown in Fig.9.

For comparison, the average observed global temperature anomalies from a 1981-2010 base period are shown in Fig.10, and the average of these anomalies is compared to the similar anomalies in the NCEP-DOE Reanalysis 2 and the CFSv2 and ECMWFv4 forecasts in Fig. 11; the predicted values are the averages of the forecasts for January and July made with leads of 1, 2, and 3 months. The range between 2000 and 2009 is approximately 0.2 C and thus is consistent with the results shown for observations in Fig 10. The differences between the observed global temperature anomalies and those of the reanalysis and the two models are shown in Fig. 12, with the CFSv2 global temperatures increasing more rapidly and the ECMWF less rapidly than the observations. The NCEP Reanalysis and the CFSv2 have similar trends even though the two series often depart from each other markedly. The NOAA Climate Prediction Center has identified and examined a number of issues related to trends in reanalyses and forecasts (e.g., Zhang, Kumar, and Wang, 2012; Xue et al., 2013).

The documentation for the two models indicates that the concentration of radiatively active gases is increased with time in an attempt to model recent trends. Thus the globally average temperature has an intentionally induced temporal trend that, as shown by Table 12 and by Fig 11, is too aggressive relative to the reanalysis for the CFSv2 and not aggressive enough for the ECMWFv4 model.

The standard statistical advice is to separate long-term trends and short-term variations and treat them independently. It seems that the seasonal modeling community is not following that advice, with the consequence that predicted probability distributions for temperature markedly favor warmth over cool relative to the verification observations. It may better serve the users of the computer seasonal forecasts to attempt to maintain a stationary state for the computations by removing long-term trends and allowing the users or their providers to make post-computational adjustments using climatologies of various lengths to fit their needs.

An increase of the predicted global temperature with decreasing lead is another unexpected bias in the seasonal forecasts, as shown in Figs 13 and 14. For example, the predicted winter global temperature increases some 0.2-0.4C between the July and December forecasts for January in the two models. These trends are summarized in Fig 15, which shows a range in the difference from average of about 0.4 C for the ECMWFv4 model and about 0.2 C for the CFSv2.

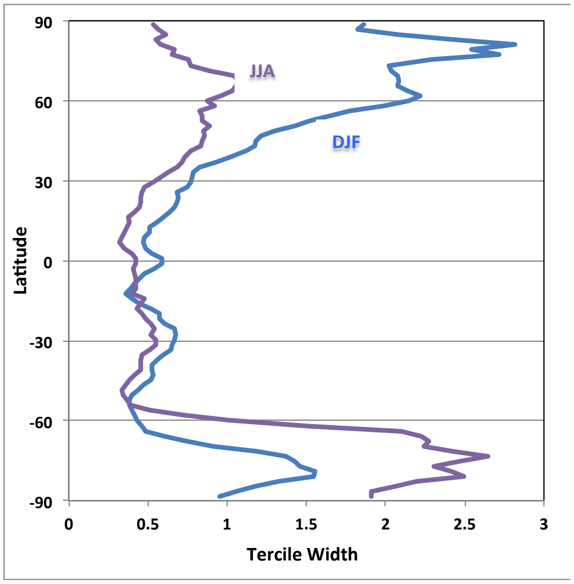


Fig. 9 Width of the near normal tercile for zonally averaged surface temperature, 1981-2010 in the NCEP-DOE Reanalysis-2 cited earlier.

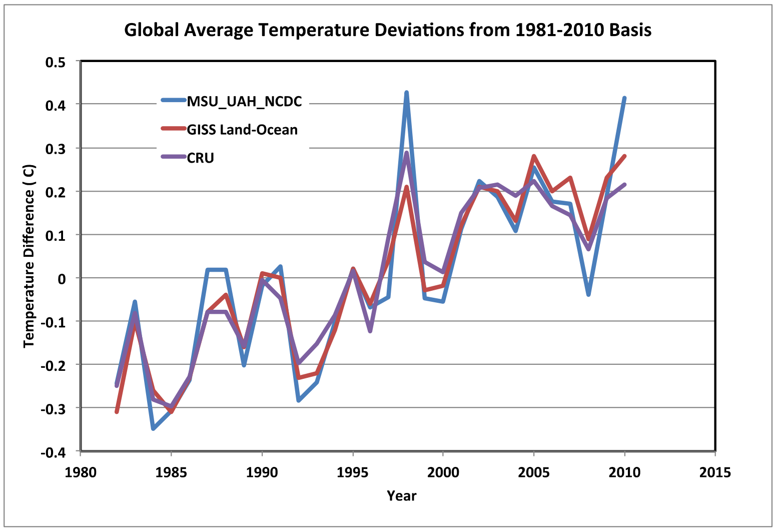


Fig. 10 Globally averaged surface temperature anomalies relative to a basis 1981-2010. The three data sets are the University of Alabama-Huntsville Microwave Sounding Unit (MSU UAH) temperatures archived at the National Climate Data Center (NCDC) and the land-ocean temperature records compiled by the NASA Goddard Institute of Space Studies (GISS) and the Climate Research Unit (CRU) of the University of East Anglia.

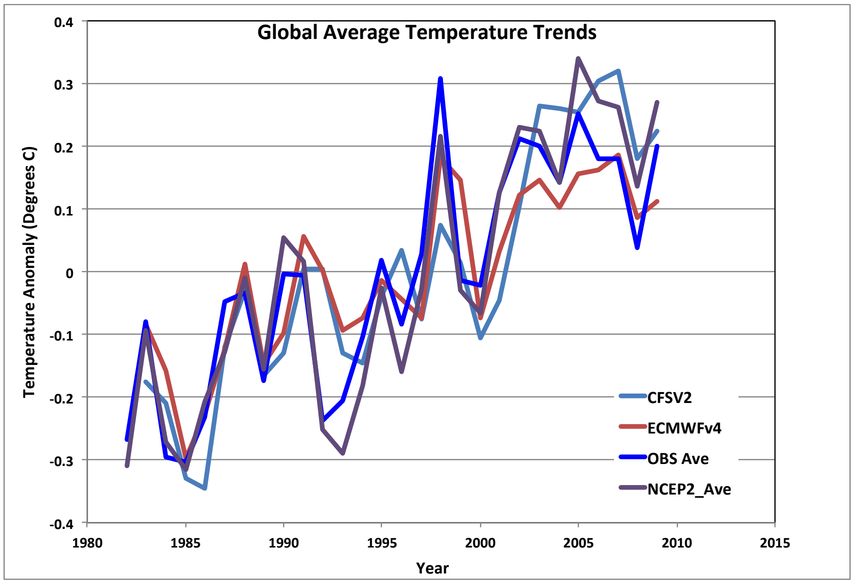


Fig. 11 Comparison of observed, NCEP-DOE Reanalysis 2, and predicted global surface temperature averages, 1982-2010

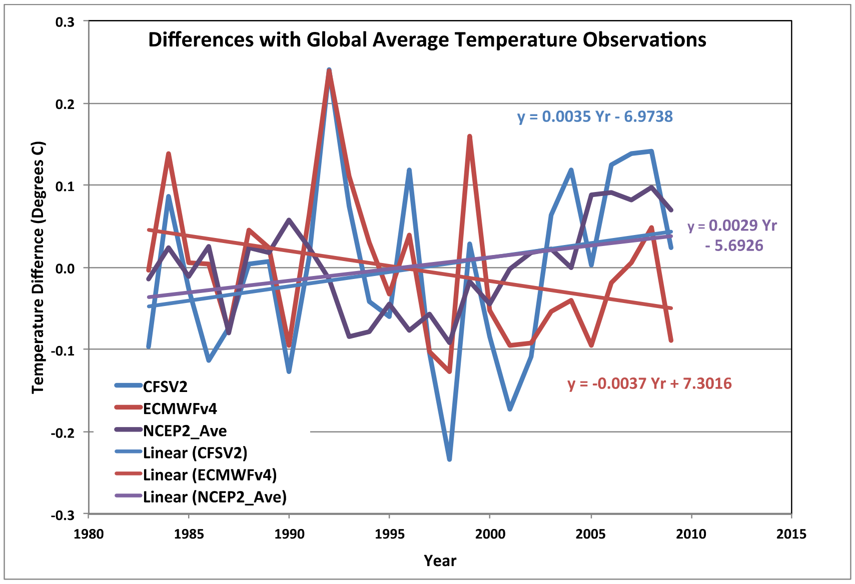


Fig. 12 Differences between the observed global annual average surface temperature and the NCEP-DOE Reanalysis 2 and the two seasonal model predictions.

Table 12 Comparison of the distributions of the NCEP-DOE Reanalysis 2 verification data and the WCS MME forecasts for surface temperature, 2000-2009 with the verification climatology (VerC). Fractions for forecasts and reanalysis verification data in percent.

	October -> Winter (DJF)			April->Summer (JJA)		
	Below Normal	Near Normal	Above Normal	Below Normal	Near Normal	Above Normal
Global Forecasts	21	24	56	19	26	56
Global VerC	26	31	43	21	33	46
Ratio	**0.80**	**0.76**	**1.30**	**0.91**	**0.77**	**1.21**
North America Forecasts	27	19	54	19	21	60
North America VerC	29	33	38	29	34	36
Ratio	**0.92**	**0.58**	**1.42**	**0.65**	**0.62**	**1.64**
Europe Forecasts	10	22	68	5	19	76
Europe VerC	28	33	40	19	35	46
Ratio	**0.37**	**0.68**	**1.71**	**0.26**	**0.53**	**1.67**
Tropical Pacific Forecasts	14	31	55	20	19	61
Tropical Pacific VerC	32	32	36	27	35	38
Ratio	**0.43**	**0.98**	**1.52**	**0.74**	**0.55**	**1.61**

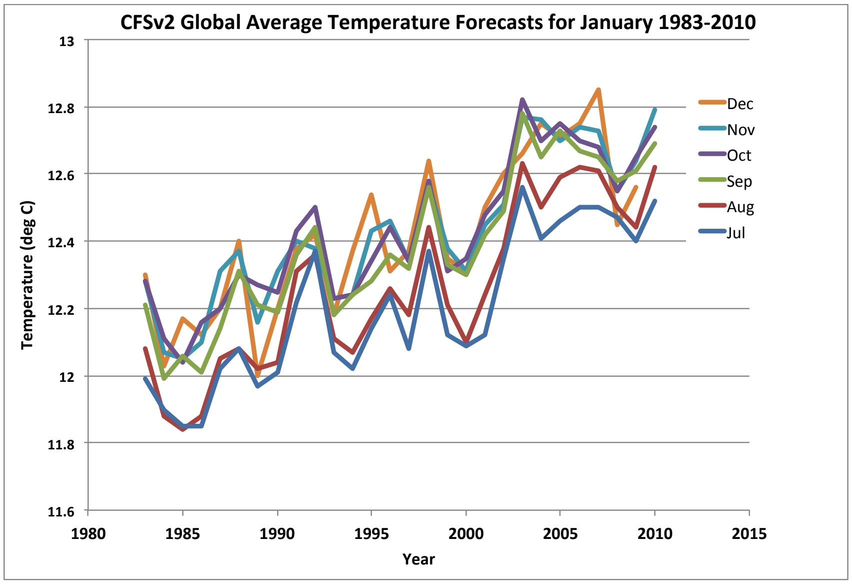


Fig. 13 Variation of CFSv2 global average temperature forecasts for January 1983-2010 as issued by the NWS for July to December.

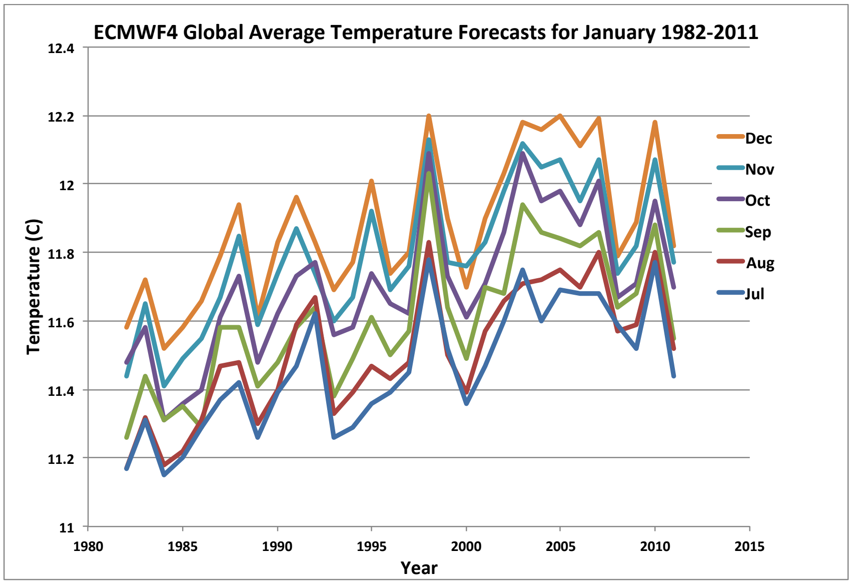


Fig. 14 Variation of the ECMWFv4 global average temperature forecasts for January 1982-2011, as issued by ECMWF for July to November.

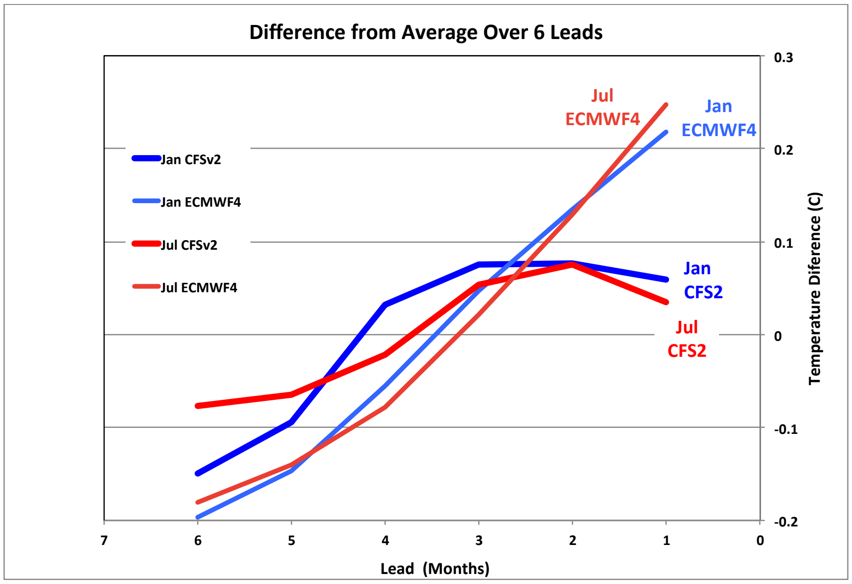


Fig. 15 Rate of increase of surface temperature in the CFSv2 and ECMWFv4 seasonal models as a function of lead.

8. Conclusion

The current versions of the CFS and ECMWF seasonal prediction models produce probability forecasts that have sufficient skill, as measured by success ratios and fractions correct, to provide meaningful assistance in managing the effects of seasonal variability. To illustrate this point as simply as possible, we showed that the WCS MME forecasts during the 10-year period considered here would have produced fairly handsome returns for investments in a hypothetical weather derivative whenever the predicted probability of a tercile exceeded 40 percent.

The WCS multi-model ensemble forecasts created with a Bayesian combination of the two models are generally more skillful and are reliable, in the sense that predicted probabilities of below, near, and above normal seasonal temperatures match the observed verification frequencies reasonably well.

The intraseasonal forecasts in the range of two-to-four week leads also provide useful skill and reliability when calibrated with an appropriate historical record.

However, there are apparent problems with long-term trends and with a strong tendency for the predicted temperature to increase as the lead decreases over a range from six to one months. It seems likely that some of this difficulty arises from an attempt to model climate change by increasing the concentrations of radiatively active gases in the models. Improved results might be obtained by attempting to maintain a statistically stationary environment for the model computations while adjusting for trends in the initial and verification observations with external statistical procedures. It may be possible to improve seasonal forecasts significantly if the issues involved in managing trends can be resolved satisfactorily.

Acknowledgments. The research reported here was supported in part by Small Business Innovation Research (SBIR) Phase I and Phase II grants from the National Oceanic and Atmospheric Administration (NOAA). Mark S. Roulston, now with Winton Capital in Oxford, introduced us to the Gaussian comb and Bayesian calibration method when we started calibrating seasonal forecasts some years ago. Comments by Roberto Buizza of the European Centre for Medium-Range Weather Forecasts (ECMWF) and an anonymous reviewer led to improvements in the article.

References

Climate Prediction Center (2011a) Toward a National Multi-Model Ensemble (NMME) System for Operational Intra-Seasonal to Interannual (ISI) Climate Forecasts, http://www.cpc.ncep.noaa.gov/products/ctb/MMEWhitePaperCPO_revised.pdf

Climate Prediction Center (2011b) NCEP-DOE AMIP-II Reanalysis, http://www.cpc.ncep.noaa.gov/products/wesley/reanalysis2/index.html

Dempster AP, NM Laird, DB Rubin (1977) Maximum likelihood from incomplete data via the EM algorithm. *J Royal Stat Soc* 39B:1-39

Doblas-Reyes FJ, R Hagedorn, TN Palmer (2005) The rationale behind multi-model ensembles in seasonal forecasting—II. Calibration and combination. *Tellus* 57A:234-252

ECMWF (2012) The EUROSIP Seasonal Forecasting System, http://www.ecmwf.int/products/forecasts/seasonal/documentation/eurosip/index.html

Johnson C, N Bowler (2009) On the Reliability and Calibration of Ensemble Forecasts, *Mon Wea Rev 137*:1717-1720.

Kim HM, PJ Webster, JA Curry (2012), Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter, *Clim Dyn* 39:2957-2973

Livezey RE, MM Timofeyeva (2008) The First Decade of Long-Lead U.S. Seasonal Forecasts: Insights from a Skill Analysis, *Bull Amer Meteorol Soc* 89:843-853.

Molteni F et al (2011) The new ECMWF seasonal forecast system (System 4), ECMWF Technical Memorandum 656, http://www.ecmwf.int/publications/library/do/references/list/14

Palmer TN (2004), Development of a European multi-model ensemble system for seasonal to inter-seasonal annual prediction (DEMETER), *Bull Amer Meteorol Soc* 85:853-872

Raftery AE, T Gneiting, F Balabdaoui, M Polakowski 2005 Using Bayesian model averaging to calibrate forecast ensembles, *Mon Wea Rev*, 133:1155-1174

Saha S et al (2010) The NCEP Climate Forecast System Reanalysis, *Bull Amer Meteorol Soc* 91: 1015-1057

Saha S et al (2013) The NCEP Climate Forecast System Version 2, submitted to *J Clim*, manuscript available at http://cfs.ncep.noaa.gov/cfsv2.info/CFSv2_paper.pdf

Taylor JW, R Buizza (2006), Density forecasting for weather derivative pricing, *Int. J. Forecast* 22: 29-42.

Van den Dool H (2007) Empirical Methods in Short-Term Climate Prediction. Oxford University Press, Oxford

Wilks DS (2006) Statistical Methods in the Atmospheric Sciences, 2nd Edition, Academic Press, Amsterdam

Xue Y, M Chen, A Kumar, Z Hu, W. Wang, 2013: Prediction Skill and Bias of Tropical Pacific Sea Surface Temperatures in the NCEP Climate Forecast System Version 2. J Climate (accepted)

Zhang, L, A Kumar, W Wang, 2012: Influence of changes in observations on precipitation: A case study for the Climate Forecast System Reanalysis (CFSR), *J Geophys Res,117*, D08105, D08105, doi:10.1029/2011JD017347