Forecast Comparison Based on Random Walks

TIMOTHY DELSOLE

George Mason University, Fairfax, Virginia, and Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland

MICHAEL K. TIPPETT

Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York, and Center of Excellence for Climate Change Research, Department of Meteorology, King Abdulaziz University, Jeddah, Saudi Arabia

(Manuscript received 8 June 2015, in final form 13 August 2015)

ABSTRACT

This paper proposes a procedure based on random walks for testing and visualizing differences in forecast skill. The test is formally equivalent to the sign test and has numerous attractive statistical properties, including being independent of distributional assumptions about the forecast errors and being applicable to a wide class of measures of forecast quality. While the test is best suited for independent outcomes, it provides useful information even when serial correlation exists. The procedure is applied to deterministic ENSO forecasts from the North American Multimodel Ensemble and yields several revealing results, including 1) the Canadian models are the most skillful dynamical models, even when compared to the multimodel mean; 2) a regression model is significantly more skillful than all but one dynamical model (to which it is equally skillful); and 3) in some cases, there are significant differences in skill between ensemble members from the same model, potentially reflecting differences in initialization. The method requires only a few years of data to detect significant differences in the skill of models with known errors/biases, suggesting that the procedure may be useful for model development and monitoring of real-time forecasts.

1. Introduction

This paper is concerned with comparing the skill of two forecasts. One of the most elegant methods of comparing skill is the *sign test*. The procedure is simple: given a criterion for selecting the most skillful forecast of a single event, count the number of times that forecast A has more skill than forecast B. If the count is ''large,'' then forecast A is more skillful than forecast B, whereas if the count is ''small'' then A is less skillful than B. To define large and small, the counts are compared to the frequency assuming that forecast A has a 50% probability on any given event of being more skillful than forecast B. A forecaster can then simply count the number of times a forecast is more skillful than another, and reject the ''equal skill'' hypothesis if the probability of obtaining that count, or a more extreme count, is less than some predefined significance threshold. This procedure is equivalent to testing whether a coin is fair based on the frequency of heads.

Often the criterion for selecting the most skillful forecast is based on the difference between two skill measures. In such cases, the sign test depends only on the *sign* of this difference (hence its name), not on the *size*. Alternative measures like correlation and mean square error account for the size of the error, but testing the significance of differences is problematic because standard tests assume the measures were computed from independent samples. In practice, skill measures computed on a common period or with a common set of observations are not independent (e.g., different forecasts tend to bust for the same event). Therefore, such tests cannot properly compare forecasts, and in fact applying these tests when the skill measures are not independent leads to serious biases (DelSole and Tippett 2014). In contrast, the sign test avoids this problem partially *because* it ignores the size of the errors. Moreover, the sign test makes no distributional assumptions about the forecast errors and is valid for a wide class of criteria for selecting the most skillful forecast. In particular, the forecast errors can be

Corresponding author address: Timothy DelSole, George Mason University, 4400 University Dr., 112 Research Hall, Mail Stop 2B3, Fairfax, VA 22030.
E-mail: delsole@cola.iges.org

DOI: 10.1175/MWR-D-15-0218.1

non-Gaussian, the skill measure need not be quadratic, symmetric, or continuous, and the criterion can be based on categorical measures.

We propose further exploiting the virtues of the sign test by displaying the results as a random walk. Specifically, whenever forecast A is more skillful than forecast B, a step in the positive direction is taken, otherwise a step in the negative direction is taken. The resulting random walk displays the *time evolution* of skill differences, thereby adding information beyond the mere decision to accept or reject the null hypothesis. Despite its simplicity, the procedure gives a revealing and easy-to-understand assessment of the relative skills of different models, as we will show. The full procedure is discussed in the next section and illustrated with a multimodel forecast dataset in sections 3 and 4. We close with a summary and discussion of our results.

2. Method

We first describe the sign test, which Diebold and Mariano (1995) proposed for comparing economic forecasts. The sign test requires a criterion for deciding the most skillful forecast of a *single* event. Aside from measures like correlation, which cannot be evaluated for a single event, virtually any criterion for judging forecasts of single events is appropriate, including categorical criteria. For instance, the criterion could be defined to focus on strong high-skill cases. More exotic applications might define "event" based on multiple forecasts, such as at a few lead times. In this paper, the forecast closest to the observation is defined to be more skillful, which is equivalent to using squared error or absolute error for the criterion. Ties present no special problems for the sign test: if a criterion implies that forecasts of a single event are equally skillful, then just eliminate these events from analysis. We assume below that no ties occur (which is true for our data).

Suppose forecasts A and B are compared N times. Let K denote the number of times A is more skillful than B. A natural null hypothesis is that at each time t forecast A has 50% probability of being more skillful than forecast B. If each time step also is independent of the others (an assumption discussed in more detail shortly), then the count K should follow a binomial distribution with $p = 1/2$, in which case the probability that the count equals K is

$$p_b(K) = \frac{1}{2^N}\frac{N!}{K!(N-K)!}. \tag{1}$$

To test the null hypothesis, we compute the probability of obtaining an observed value K_0, or a more extreme value, from (1). This probability is called the p value and is given by

$$p \text{ value} = 2p_b(0) + 2p_b(1) + \cdots + 2p_b(\min[K_o, N - K_o]), \tag{2}$$

where the factor of 2 accounts for the fact that test is two tailed, since a priori we do not know which forecast is superior. We reject the null hypothesis if the p value falls below a prescribed significance level α. This procedure is called the *sign test* (Conover 1980) and has been applied for forecast verification (Diebold and Mariano 1995; DelSole and Tippett 2014). The sign test is equivalent to testing the hypothesis that the *median* measure for deciding the most skillful forecast vanishes. Also, the sign test is equivalent to testing the hypothesis that the forecast comparisons are drawn from independent Bernoulli trials with $p = 1/2$.

The critical value K_α^{binom} is the smallest value of K_o such that the p value in (2) is equal to or greater than α. Critical values for $N = 1, 2, \ldots, 45$ are shown in Table 1 for $\alpha = 5\%$. The hypothesis is rejected if $K_o < K_\alpha^{\text{binom}}$ or $K_o > N - K_\alpha^{\text{binom}}$. For large N, the binomial distribution has an approximately Gaussian distribution with mean $N/2$ and variance $N/4$. If one also invokes a standard "continuity correction" to adjust for the use of a continuous distribution to approximate a discrete distribution (Rosner 2000), then the critical value can be approximated as

$$K_\alpha^{\text{norm}} = \lceil N/2 - z_{\alpha/2}\sqrt{N/4} - 1/2 \rceil, \tag{3}$$

where z_α is the value for which a standardized Gaussian is exceeded with probability α, and $\lceil x \rceil$ denotes the smallest integer not less than x. The critical values derived from the Gaussian approximation, shown in Table 1, are mostly identical to the critical values derived from the exact binomial distribution, even for small N, and hence will be used in the remainder of this paper.

We propose expressing the counts in terms of a random walk. Specifically, whenever A is more skillful than B, a step in the positive direction is taken; otherwise, a step in the negative direction is taken (ties are assumed to never occur). Accordingly, there are K steps in the positive direction and $N - K$ steps in the negative direction, so the net distance traveled by the random walk is

$$d_N = K - (N - K) = 2K - N. \tag{4}$$

Using (3), an approximate confidence can be computed from (4) as

TABLE 1. The exact ($K_{0.05}^{\text{binom}}$) and Gaussian approximated ($K_{0.05}^{\text{norm}}$) lower critical value of the number of trials, out of N, a given forecast can exceed the skill of another forecast, in order to reject the hypothesis of a binomial distribution with $p = 1/2$ at the 5% significance level.

N	$K_{0.05}^{\text{norm}}$	$K_{0.05}^{\text{binom}}$	N	$K_{0.05}^{\text{norm}}$	$K_{0.05}^{\text{binom}}$	N	$K_{0.05}^{\text{norm}}$	$K_{0.05}^{\text{binom}}$	N	$K_{0.05}^{\text{norm}}$	$K_{0.05}^{\text{binom}}$
1	0	0	16	4	4	31	10	10	46	16	16
2	0	0	17	5	4	32	10	10	47	17	17
3	0	0	18	5	5	33	11	11	48	17	17
4	0	0	19	5	5	34	11	11	49	18	18
5	0	0	20	6	6	35	12	12	50	18	18
6	1	1	21	6	6	36	12	12	51	19	19
7	1	1	22	6	6	37	13	13	52	19	19
8	1	1	23	7	7	38	13	13	53	19	19
9	2	2	24	7	7	39	13	13	54	20	20
10	2	2	25	8	8	40	14	14	55	20	20
11	2	2	26	8	8	41	14	14	56	21	21
12	3	3	27	8	8	42	15	15	57	21	21
13	3	3	28	9	9	43	15	15	58	22	22
14	3	3	29	9	9	44	16	15	59	22	22
15	4	4	30	10	10	45	16	16	60	22	22

$$(1-\alpha)100\%\text{ confidence interval for } d_N \approx z_{\alpha/2}(-\sqrt{N}, \sqrt{N}). \quad (5)$$

The exact upper limit of d_N differs from $z_{\alpha/2}\sqrt{N}$ by less than 1.09 for $N \leq 1000$.

The above test assumes that the forecasts are independent. In weather and seasonal prediction, forecast errors often are correlated between consecutive events. Diebold and Mariano (1995) review alternative tests of skill differences that account for serial correlation, including a generalization of the t test that is popular in the economics literature. For the sign test, however, serial correlation is problematic because large excursions in the counts are more frequent than those from an independent Bernoulli process. We have explored various methods of dealing with serial correlation, including skipping across a fixed number of months, using residuals of autoregressive model fits, and correcting the p value using Bonferroni methods. Unfortunately, these methods do not fully account for serial correlation (e.g., forecast errors remain correlated even after 12 months). Also, these methods lead to conclusions that depend on the method used to remove serial correlation and on the models being compared, thereby obscuring the final conclusion. Instead, we apply the test regardless of serial correlation. Technically, then, the test is for the entire null hypothesis of *independent* Bernoulli trials. We argue that rejecting the hypothesis of independent Bernoulli trials is informative even if due to serial correlation, because if yesterday's forecast A was better than B, then tomorrow's forecast A ought to be better than B. Therefore, hedging toward forecast A ought to be better than simply averaging forecasts A and B. In this way, a significant difference relative to independent Bernoulli trials can be useful for improving forecasts, even if the difference is due to serial correlation. Exactly how much to hedge is a question that requires separate study. The test provides an objective basis for deciding whether such study is warranted.

Note that the above test is free of distributional assumptions regarding the error and is applicable to a wide class of criteria for selecting the most skillful forecast. This is in contrast to correlation skill or mean square error, whose significance tests often are based on the Gaussian assumption.

3. Data

We illustrate the skill comparison test using hindcasts of monthly mean Niño-3.4 from the North American Multimodel Ensemble (NMME). The NMME, reviewed by Kirtman et al. (2014), consists of at least 9-month hindcasts by state-of-the-art coupled atmosphere–ocean models from the following centers: the National Centers for Environmental Prediction (NCEP-CFSv1 and NCEP-CFSv2), the Canadian Centre for Climate Modeling and Analysis (CMC1-CanCM3 and CMC2-CanCM4), the Geophysical Fluid Dynamics Laboratory (GFDL-CM2p1-aer04, GFDL-CM2p5-FLOR-A06, and GFDL-CM2p5-FLOR-B01), the International Research Institute for Climate and Society (IRI-ECHAM4p5-Anomaly and IRI-ECHAM4p5-Direct), the National Aeronautics and Space Administration (NASA-GMAO-062012), and a joint collaboration between the Center for Ocean–Land–Atmosphere Studies, University of Miami, and

TABLE 2. List of NMME models and relevant details. Entries under "ensemble generation" summarize our assessment of whether the initialization system (which differs for each model) is exchangeable, or if not, the potential cause for the lack of exchangeability. (Expansions of acronyms are available online at http://www.ametsoc.org/PubsAcronymList.)

Full model name	Shortened model name	First real-time forecast	Status	Ensemble generation
NCEP-CFSv1	CFSv1	2011	Retired	Lagged
NCEP-CFSv2	CFSv2	2011	Active	Lagged
CMC1-CanCM3	CanCM3	2011	Active	Exchangeable
CMC2-CanCM4	CanCM4	2011	Active	Exchangeable
GFDL-CM2p1-aer04	CM2p1-aer04	2011	Active	Exchangeable
GFDL-CM2p5-FLOR-A06	FLOR-A	2014	Active	Exchangeable
GFDL-CM2p5-FLOR-B01	FLOR-B	2014	Active	Exchangeable
IRI-ECHAM4p5-Anomaly	IRI-A	2011	Retired	Exchangeable
IRI-ECHAM4p5-Direct	IRI-D	2011	Retired	Exchangeable
NASA-GMAO-062012	NASA	2011	Active	Some members lagged
COLA-RSMAS-CCSM3	CCSM3	2011	Active	Fixed atmospheric ICs
COLA-RSMAS-CCSM4	CCSM4	2014	Active	Lagged atmosphere

the National Center for Atmospheric Research (COLA-RSMAS-CCSM3 and COLA-RSMAS-CCSM4).

Relevant details of the above models are given in Table 2. Real-time forecasts began in August 2011. Three of the above models began real-time forecasts in 2014: GFDL-CM2p5-FLOR-A06, GFDL-CM2p5-FLOR-B01, and COLA-RSMAS-CCSM4. Three NMME models have been retired: NCEP-CFSv1, IRI-ECHAM4p5-Anomaly, and IRI-ECHAM4p5-Direct.

The CFSv2 hindcasts have an apparent discontinuity across 1999, presumably due to the introduction of certain satellite data into the assimilation system in October 1998 (Kumar et al. 2012; Barnston and Tippett 2013; Saha et al. 2014). Different bias corrections will be considered in the next section, but all of them will avoid computing climatologies over periods that cross 1999.

The validation data used in this study are the NOAA Optimum Interpolation Sea Surface Temperature (OISST) version 2 (Reynolds et al. 2002). The variable investigated is the Niño-3.4 index, which is the area weighted sea surface temperature within the region bounded by 5°S– 5°N, 120°–170°W. (The OISST version of this index is available from the Climate Prediction Center website at http://www.cpc.ncep.noaa.gov/data/indices/sstoi.indices.)

4. Results

For convenience, we use the term "forecast" to refer to both ensemble mean hindcasts and forecasts. To eliminate differences due to different ensemble sizes, the same number of ensemble members per model is analyzed. Specifically, we select the last six ensemble members of CFSv1 and CFSv2 (because these members were initialized closest to the target date, as discussed shortly) and the first six members of all other models. This selection results in over 200 000 forecasts (e.g., 13 models, 8 lead months, 12 months per year, 33 years, and 6 ensemble members corresponds to 247 104 forecasts). To reduce the analysis to a manageable size, we focus only on ensemble mean forecasts of the Niño-3.4 index at 2.5-month lead (e.g., for a March target, the model is initialized using observations no later than mid-January). The most skillful forecast is defined to be the ensemble mean forecast closest to the observed monthly mean Niño-3.4 index. Equivalently, the most skillful forecast has the least squared error or least absolute error between forecasted and observed monthly mean Niño-3.4 index. We choose the significance level $\alpha = 5\%$, in which case $z_{\alpha/2} \approx 1.96$.

a. *Exchangeability*

As a novel application of our method, we first apply it to test differences in skill among ensemble members *from the same model*. Ensemble members are intended to be exchangeable, in the sense that the statistical properties of the ensemble should be invariant to permutations of the member labels. For instance, some models use members generated by randomly perturbing the same initial state. If the random numbers are drawn independently from the same distribution, no statistical feature could exist to discriminate between members, hence, the members are exchangeable. Similarly, the GFDL models use ocean ensemble data assimilation schemes, which generate exchangeable members (the models also use states from atmospheric models that are exchangeable). In contrast, some models use a *lagged* ensemble in which different members correspond to different start dates. Lagged ensembles might be distinguishable because members initialized farther from the target may have more skill than those initialized closer to the target. In addition to a lagged ensemble, the NASA model also includes members generated by special perturbation techniques, such as breeding methods, which might be more or less skillful than forecasts

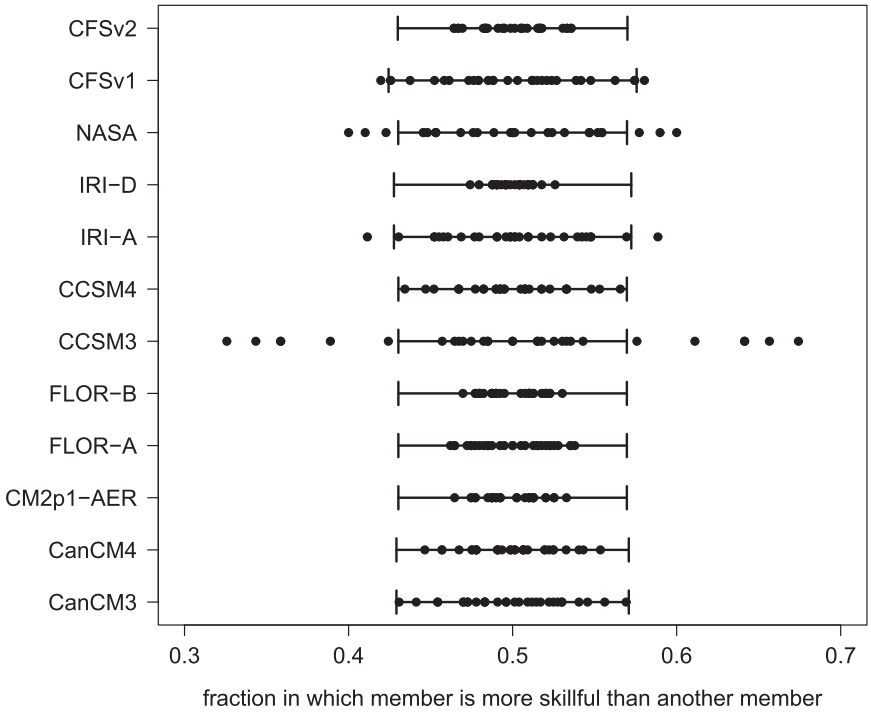

FIG. 1. Fraction of forecasts from a given model in which one ensemble member is more skillful than another member, for all possible pairs of ensemble members, for 2.5-month lead during 1982–2014 (dots), and the 95% confidence interval for independent trials of a Bernoulli process with $p = 1/2$ (error bar).

generated by other initialization techniques. Different institutions use different initialization schemes on different components of the coupled atmosphere–ocean–land–ice system, some of which are quite complex. Therefore, it is difficult to summarize these schemes in this paper. In Table 2 we have summarized our assessment of whether each initialization scheme, as documented in the literature, is exchangeable, and if not, speculated as to the reasons for differences in skill between members.

Exchangeability can be tested partly by comparing the skill between two ensemble members from the same model: if one member is more skillful than another, then obviously the two members are not exchangeable. The result of testing all possible pairs of ensemble members for all available months and years is shown in Fig. 1. The 95% confidence interval for independent Bernoulli trials is indicated by the error bars. For most models, the results are consistent with that expected for independent Bernoulli trials. Differences in skill between ensemble members are evident in CFSv1, NASA, IRI-A, and CCSM3. CFSv1 and NASA use lagged ensembles, which are not strictly exchangeable (e.g., members initialized closer to the target are likely to be more skillful), so we presume the test has detected differences due to using lagged ensembles.

Almost half of the comparisons from CCSM3 are more skillful than expected. Closer inspection reveals that ensemble member "2" has significantly less skill than other members. The reason for this difference is unclear. For CCSM3, the ocean state is identical for all members initialized in the same month, but the atmosphere, land, and ice are drawn from different years in a long control simulation (B. Kirtman 2015, personal communication). Importantly, the initial atmosphere–land–ice state for a given month and ensemble member is identical across years (e.g., the atmosphere–land–ice state in the first ensemble member for January 1982 is identical with that of the first ensemble member for January 1983, January 1984, etc.). Strictly speaking, then, the initial conditions for the atmosphere–land–ice state are not statistically exchangeable—in particular, the initial state for ensemble member 2 may have a large bias relative to other members.

Interestingly, CFSv2 also uses a lagged ensemble, but no differences in skill among ensemble members were detected. This might be because the lagged ensemble for CFSv2 is much more closely spaced than for CFSv1. For example, for a mid-June release, our particular CFSv2 ensemble has four members initialized on 5 June and two members initialized on 31 May (Saha et al. 2014). In contrast, CFSv1 has one member is initialized on 21 May and others initialized on 30 and 31 May and 1, 2, and 3 June (Saha et al. 2006). Thus, the most extreme time

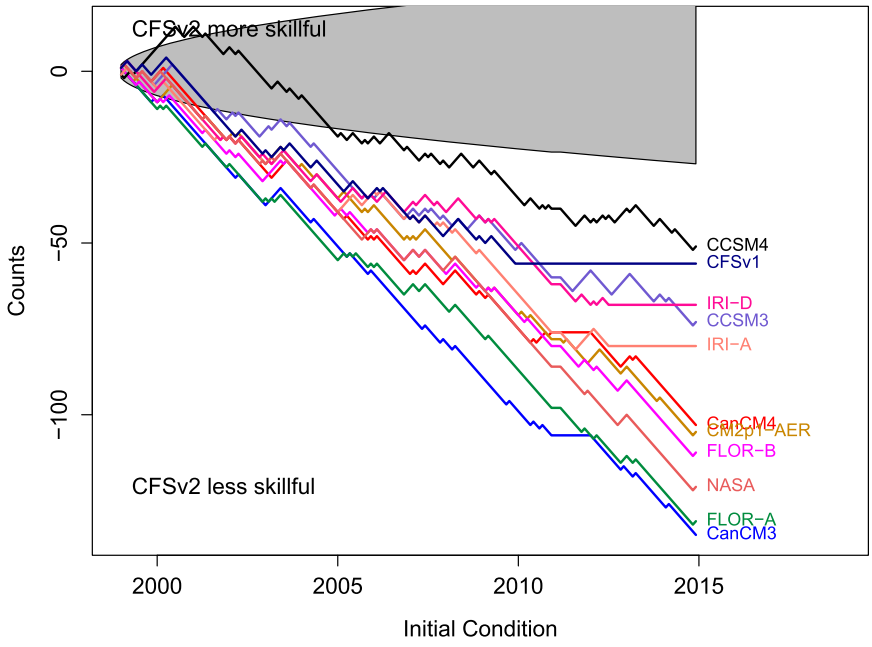

FIG. 2. Comparison of monthly mean forecasts of Niño-3.4 at 2.5-month lead between CFSv2 and other models in the NMME. The mean 1982–98 error is removed from each model. The count increases by 1 when the squared error of CFSv2 is less than that of another model, and decreases by 1 otherwise. The count is accumulated forward in time for each model separately, over all initial months and years (for a fixed lead time), thereby tracing out a random walk. The shaded area indicates the range of counts that would be obtained 95% of the time under independent Bernoulli trials for $p = 1/2$. A random walk extending above the shaded area indicates that CFSv2 forecasts are closer to observations significantly more often than expected for independent Bernoulli trials (i.e., the CFSv2 is more skillful than the model). Conversely, a random walk extending below the shaded area indicates that CFSv2 forecasts are closer to observations significantly less often than expected for independent Bernoulli trials (i.e., the CFSv2 is less skillful than the model).

separation for our CFSv2 six-member ensemble is 6 days, while that for CFSv1 is 13 days. Since the time separation is larger for CFSv1 than for CFSv2, the CFSv1 members are likely to have larger differences in skill.

Differences in skill are detected for members from IRI-A, but the source of these differences are unclear. Note that 5% of the cases on average would lie outside the computed limits even if exchangeability were true.

b. Bias correction based on the 1982–98 mean error

We now consider bias-corrected forecasts. Ideally, the same data should not be used to estimate bias and to compare skill simultaneously, since the sign test does not account for bias correction, moreover such correction is unrealistic in real-time forecasting because future data are not available for estimating bias. A straightforward way to circumvent this problem is to estimate the bias correction parameters using data that are separate from that used to test skill differences. Accordingly, we estimate the mean forecast error for each calendar month using hindcasts whose verifications lie within the period 1982–98 inclusive, and then subtract this error from the appropriate forecast after this period. The comparisons begin with hindcasts initialized on January 1999. The result of comparing CFSv2 with other models, shown in Fig. 2, reveals that all models are significantly more skillful than CFSv2 after only a few years of comparisons. The poor skill of CFSv2 relative to other models has been attributed to a discontinuity in climatology due to the introduction of ATOVS satellite data into the assimilation system in October 1998, as discussed in Kumar et al. (2012), Barnston and Tippett (2013), and Saha et al. (2014). Thus, the result is not unanticipated, but the rapidity and decisiveness of detection of an abrupt difference in skill is noteworthy.

The above illustration may seem contrived because the transition year 1999 was purposely avoided when estimating the bias correction. In practice, a real forecaster would not ordinarily know the transition year of an abrupt change in skill. Fortunately, changes in skill can be recognized by changes in the *average slope* of a random walk. To illustrate this fact, we show in Fig. 3 the

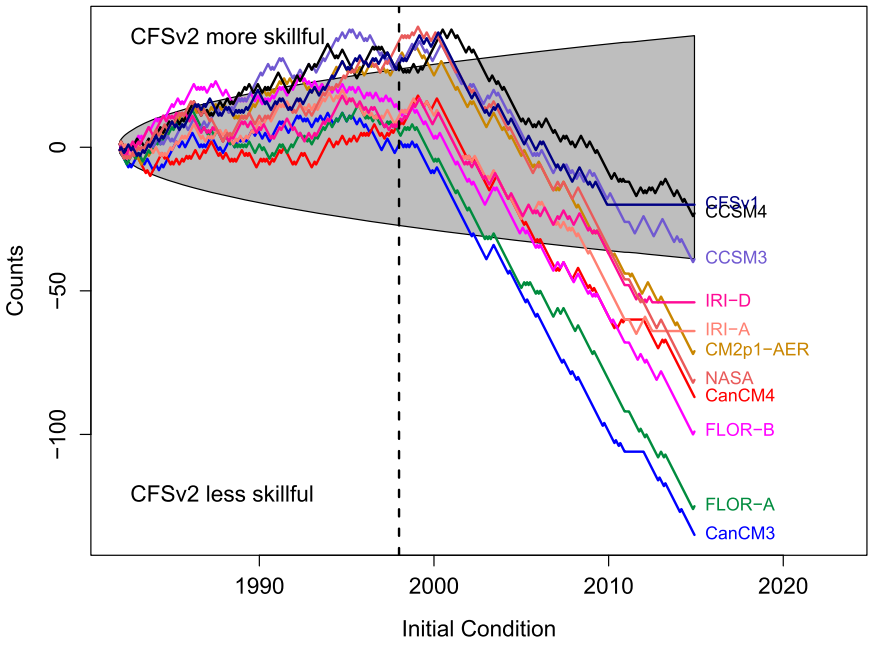

FIG. 3. Comparison of 2.5-month lead forecasts of monthly mean Niño-3.4 from CFSv2, as in Fig. 2, but the comparisons are initialized in 1982 (all anomalies are computed with respect to the 1982–98 mean).

result of comparing CFSv2 with other models, except that the random walk is initialized in 1982. The figure clearly shows a distinct change in average slope beginning around 2000. This result illustrates how random walks could be used in real time to monitor and detect abrupt changes in skill without knowledge of the transition year.

Comparisons based on the ''retired'' models IRI-A, IRI-D, and CFSv1 are shown in Fig. 4. Forecasts for these models were unavailable at the end of the period. In such cases, the ''count'' is kept constant, which produces a random walk that is ''flat'' at the end of the period. Incidentally, this flattening procedure can be applied when ties occur in forecast skill, although such ties never occur in our data. The figure shows that, among other things, no significant difference in skill is detected between IRI-A and IRI-D. Thus, at least for ENSO forecasts from these two models, no detectable difference in skill can be detected between anomaly coupled and fully coupled models, provided a bias correction is applied. The fact that all three models are more skillful than CFSv2 (using anomalies with respect to 1982–98 climatology) has already been indicated in Fig. 2.

The comparison between other models not considered yet are shown in Fig. 5 (still using a bias correction derived from 1982–98). The figure reveals that the Canadian models are significantly more skillful than other models. CFSv2 is significantly less skillful than all other models, as discussed earlier. CCSM4 and CCSM3 are the next least skillful models. FLOR-A and FLOR-B are either comparable to, or significantly more skillful than, all other models except for the Canadian models. Moreover, no significant difference in skill is detected between FLOR-A and FLOR-B. Also, FLOR-A and FLOR-B represent an improvement over the previous version of the GFDL model CM2p1-AER. The NASA model is either comparable to, or significantly more skillful than, other models except for the Canadian models.

c. *Multimodel mean*

It is interesting to compare NMME forecasts with the multimodel mean. To compute the multimodel mean, we omit the retired models IRI-A, IRI-D, and CFSv1. Since we continue removing the mean bias estimated during 1982–98, we omit CFSv2 due to the discontinuity around 1999. Note that our test can compare the multimodel mean with an individual model even if that model is contained in the multimodel mean. The resulting comparison, shown in Fig. 6, reveals that the multimodel mean is either comparable to, or more skillful than, other models except for the Canadian models. These and subsequent results are unchanged if CCSM4 is omitted from the multimodel mean. Thus, although CCSM4 is significantly less skillful than all other models (see Fig. 5), it does not significantly alter the skill of the multimodel mean.

d. *Statistical forecasts*

Another interesting question is how the dynamical forecasts compare to purely statistical forecasts. Although there

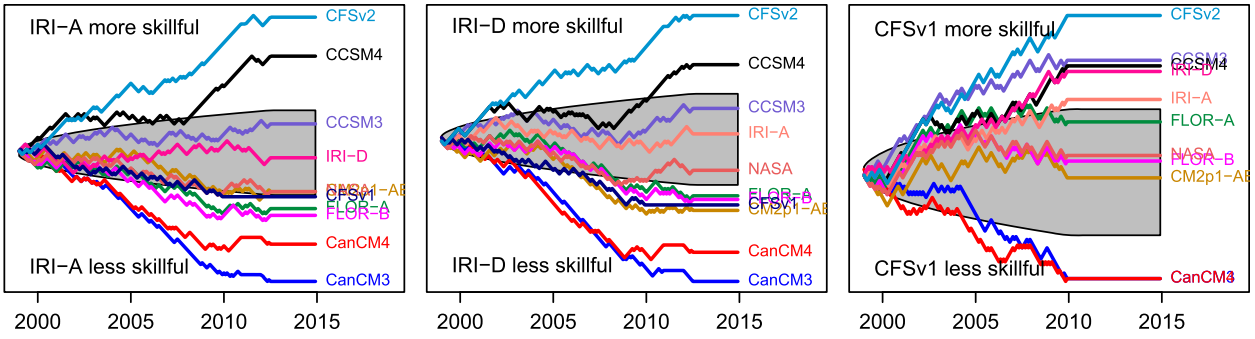

FIG. 4. Comparison of 2.5-month lead forecasts of monthly mean Niño-3.4, as in Fig. 2, but for the "retired" models IRI-A, IRI-D, and CFSv1, and with the mean 1982–98 error removed from each model.

exists a wide variety of statistical models of ENSO, here we use ordinary linear regression. Specifically, for each calendar month m, we fit the linear prediction equation:

$$T_{m+\tau} = b_{m,\tau} + a_{m,\tau} T_{m-1} + \epsilon, \tag{6}$$

where $a_{m,\tau}$ and $b_{m,\tau}$ are the slope and intercept terms for month m, respectively; τ is the lead time (i.e., 2.5 months); and ϵ is the random error. The predictor T_{m-1} has an extra 1-month lag for consistency with the NMME forecast protocol. For instance, for a 0.5-month lead forecast in November, the observed state for November is unavailable at the beginning of the forecast. Thus, a regression prediction must use the observed state for the preceding October to produce a forecast for November. The slope and intercept terms are estimated by standard linear regression. The training period for this estimation corresponds to verifications between March 1982 and December 1998 inclusive (for lead = 2.5 months). The prediction $\hat{T}_{m+\tau}$ is then computed as

$$\hat{T}_{m+\tau} = \hat{b}_{m,\tau} + \hat{a}_{m,\tau} T_{m-1}, \tag{7}$$

where $\hat{a}_{m,\tau}$ and $\hat{b}_{m,\tau}$ are the least squares estimates of the slope and intercept. The resulting comparison, shown in Fig. 7, reveals that the statistical model is significantly more skillful than all other models except two (CanCM3 and CanCM4). Note that the regression model is significantly more skillful than the multimodel mean too. This conclusion holds for all lags examined, namely, 0.5–7.5 months.

e. Bias correction based on the 1999–2010 mean error

We now compare forecasts initialized after 2010, but biased corrected using the 1999–2010 mean error. This comparison allows CFSv2 to be included (since the bias correction avoids the discontinuity across 1999). Also, many models generated genuine forecasts during this period. The linear regression model is retrained based on the 1982–2010 period (although results for the original regression model trained on 1982–98 are similar). The resulting comparisons are shown in Fig. 8. CCSM3 stands out as being marginally or significantly less skillful than most other models. CCSM4 is significantly more skillful than three other models (CCSM3, NASA, and CanCM4). CanCM3 is significantly more skillful than CanCM4 and CCSM3 during the forecast period, but otherwise has comparable skill to other models. The regression model is significantly more skillful than CCSM3, but otherwise has comparable skill to other models.

5. Summary and discussion

This paper proposed a procedure for testing differences in forecast skill that can be visualized as a random walk. The random walk is defined as follows: whenever forecast A is more skillful than forecast B, a step in the positive direction is taken, otherwise, a step in the negative direction is taken. If the distance traveled by the random walk after N steps falls outside the 2.5% and 97.5% interval of a binomial distribution with N and $p = 1/2$, which is approximately $(-2\sqrt{N}, 2\sqrt{N})$, then the hypothesis of equally skillful forecasts is rejected at the 5% significance level. The test is formally equivalent to the sign test, but the random walk representation further shows the *evolution* of skill differences. Remarkably, the test can be applied to general criteria for selecting the most skillful forecast and is independent of distributional assumptions about the forecast errors. The method also can be used to compare a multimodel mean with another model that may be included in the multimodel mean, in contrast to most standard tests.

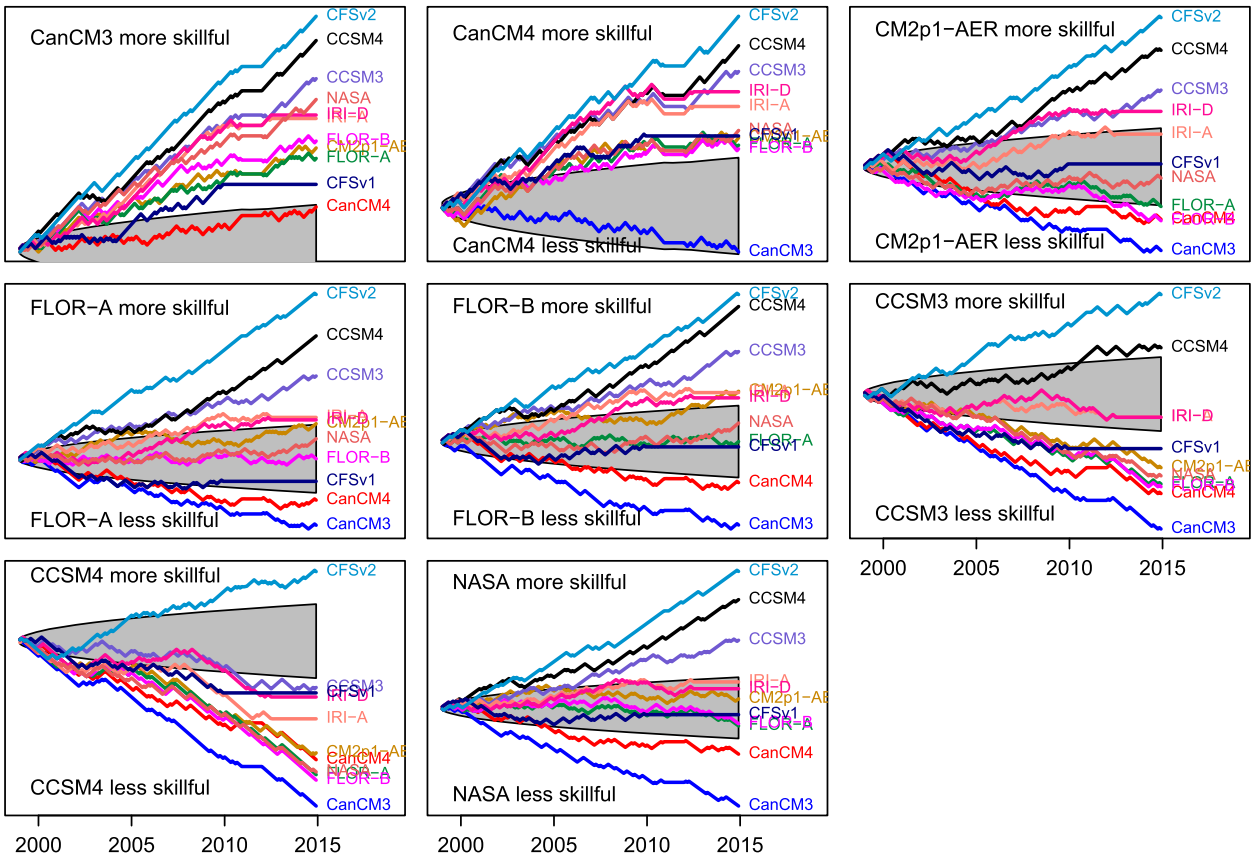

FIG. 5. Comparison of 2.5-month lead forecasts of monthly mean Niño-3.4, as in Fig. 2, but for other models not considered in previous figures. The mean 1982–98 error is removed from each model.

The above test is most suited for independent events. In practice, forecast errors tend to be serially correlated, especially when separated by short time periods (e.g., numerical weather predictions separated by a day or less). We do not recommend "correcting" for independence by adjusting degrees of freedom or applying whitening transformations, because such methods are difficult to justify and require estimating parameters using most of the data. In addition, such methods lead to conclusions that are difficult to interpret because they depend on the method used to account for serial correlation. Instead, we recommend applying the test in its pure form and clearly stating that the hypothesis being tested is *independent* Bernoulli trials with $p = 1/2$. We argue that the result is useful even if serial correlation exists. For instance, if a forecast is more skillful and this skill is persistent, then hedging toward previous more skillful forecasts can lead to better forecasts at subsequent times compared to equal weighting schemes. The test can be viewed as an objective procedure for deciding whether such hedging is warranted.

The above procedure was applied to NMME monthly mean hindcasts and forecasts of Niño-3.4 at 2.5-month lead. The procedure was able to detect the discontinuity in skill of CFSv2 after 1998. This result is not surprising in light of known errors in this model, but it is significant that the procedure was able to detect this difference in skill after only a few years of hindcasts. These results illustrate how the proposed test may be an effective tool for routine model development. Over the period 1999–2014, a possible ranking of the Niño-3.4 forecasts at 2.5-month lead is as follows:

1) CanCM3, CanCM4, linear regression model;
2) FLOR-A, FLOR-B, multimodel mean, NASA;
3) CM2p1-AER;
4) CCSM3; and
5) CCSM4.

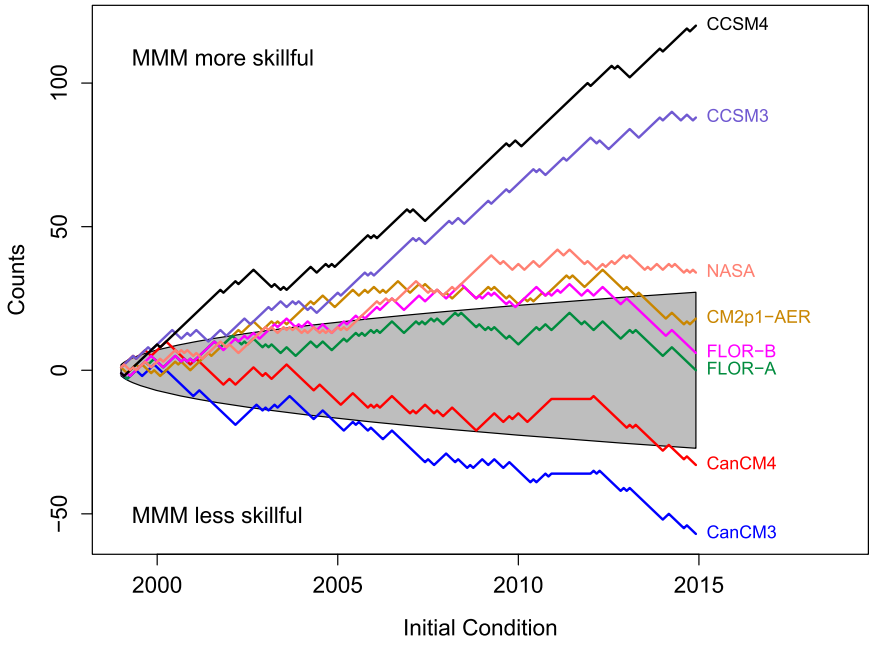

FIG. 6. Comparison of 2.5-month lead forecasts of monthly mean Niño-3.4, as in Fig. 2, but for the multimodel mean based on the models indicated in the figure. The mean 1982–98 error is removed from each model.

Technically, a unique ranking is ill defined because some pairwise comparisons are intransitive; for example, FLOR-A is comparable to NASA and NASA is comparable CM2p1-AER, but FLOR-A is significantly more skillful than CM2p1-AER. Thus, NASA could be grouped into category 2 or 3. CFSv2 is not included because the ranking is based on a bias correction estimated from the 1982–98 hindcasts, which does not produce an accurate correction for CFSv2 after 1999.

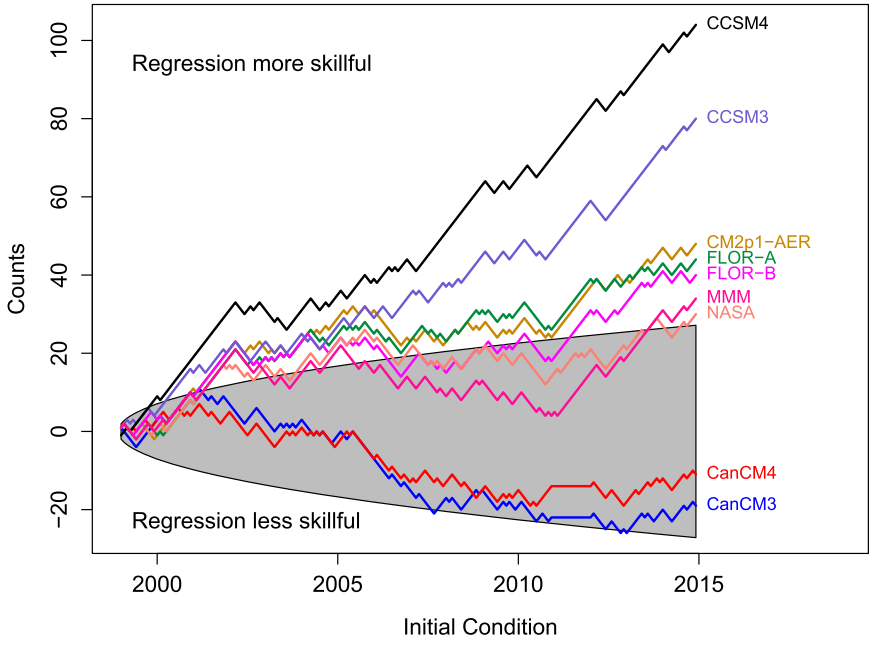

FIG. 7. Comparison of 2.5-month lead forecasts of monthly mean Niño-3.4, as in Fig. 2, but for the linear regression model based on 1982–98 training data. The mean 1982–98 error is removed from each model.

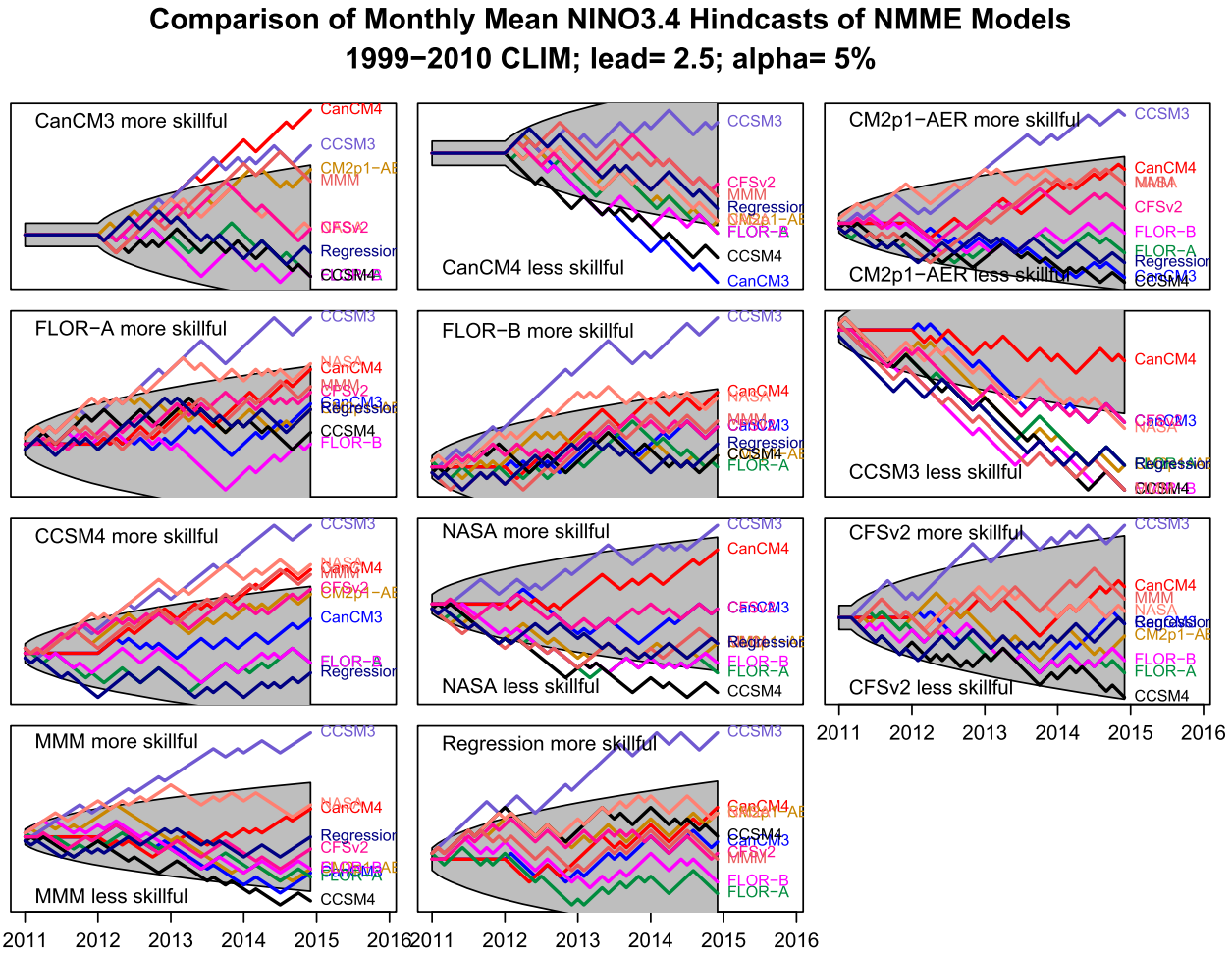

FIG. 8. Comparison of 2.5-month lead forecasts of monthly mean Niño-3.4, as in Fig. 2, but for bias correction based on 1999–2010 training data, and validated over 2011–14. The linear regression model is trained on 1982–2010 data.

Other interesting conclusions include the following. Linear regression produces more skillful predictions of monthly mean Niño-3.4 anomalies than most models in the NMME at all leads investigated (0.5–7.5 months). No significant difference in skill was detected between bias-corrected IRI-A and IRI-D forecasts, indicating little difference in skill between anomaly coupled and fully coupled prediction models. FLOR-A and FLOR-B are significantly more skillful than its predecessor CM2p1-AER, but CCSM4 is not significantly more skillful than its predecessor CCSM3 during 1999–2010, but is more skillful during 2011–14. The test also revealed significant differences in skill between ensemble members from the same model. In virtually every case for which differences in skill could be detected, the method used to generate ensemble members does not produce strictly exchangeable ensemble members. For instance, CFSv1 and NASA employ lagged ensembles, so members initialized closer to the target are likely to be more skillful and, therefore, are potentially distinguishable.

It should be recognized that the present study is limited by the fact that only a single index has been examined. Of course, a model that performs well for one index may perform poorly for another index. The Niño-3.4 index has been chosen for this study because it is one of the most important predictors of seasonal mean climate variables.

We suggest that the above method can be a very useful tool for model development. One of the biggest challenges in model development is to decide whether a particular model change improves skill. As discussed in DelSole and Tippett (2014), statistical tests based on correlation skill or mean square error are problematic. In contrast, the proposed method is completely rigorous, makes no distributional assumptions about the forecast errors, and can be applied to a wide class of criteria for selecting the most skillful forecast. Therefore, the

method can be applied even to highly non-Gaussian variables like precipitation, and can be tailored to specific performance measures of interest to modelers or forecasters. Finally, the method can detect discontinuities in skill without knowledge of when they might occur, and, therefore, can be used to decide if a significant change in skill has occurred due to a change in dynamical model, change in the quality of the initial conditions, or inadvertent errors introduced in the forecast system.

Acknowledgments. We thank Michelle L'Heureux for useful discussions and for verifying some of our results. This research was supported primarily by the National Oceanic and Atmospheric Administration, under the Climate Test Bed program (Grant NA10OAR4310264). Additional support was provided by the National Science Foundation (Grants ATM0332910, ATM0830062, and ATM0830068), the National Aeronautics and Space Administration (Grants NNG04GG46G and NNX09AN50G), and the National Oceanic and Atmospheric Administration (Grants NA04OAR4310034, NA09OAR4310058, NA05OAR4311004, NA10OAR4310210, NA10OAR4310249, and NA12OAR4310091). The views expressed herein are those of the authors and do not necessarily reflect the views of these agencies.

REFERENCES

Barnston, A. G., and M. K. Tippett, 2013: Predictions of Nino3.4 SST in CFSv1 and CFSv2: A diagnostic comparison. *Climate Dyn.*, **41**, 1615–1633, doi:10.1007/s00382-013-1845-2.

Conover, W. J., 1980: *Practical Nonparametric Statistics.* 2nd ed. Wiley-Interscience, 493 pp.

DelSole, T., and M. K. Tippett, 2014: Comparing forecast skill. *Mon. Wea. Rev.*, **142**, 4658–4678, doi:10.1175/MWR-D-14-00045.1.

Diebold, F. X., and R. S. Mariano, 1995: Comparing predictive accuracy. *J. Bus. Econ. Stat.*, **13**, 253–263.

Kirtman, B. P., and Coauthors, 2014: The North American Multimodel Ensemble: Phase-1 seasonal-to-interannual prediction; Phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, doi:10.1175/BAMS-D-12-00050.1.

Kumar, A., M. Chen, L. Zhang, W. Wang, Y. Xue, C. Wen, L. Marx, and B. Huang, 2012: An analysis of the nonstationarity in the bias of sea surface temperature forecasts for the NCEP Climate Forecast System (CFS) version 2. *Mon. Wea. Rev.*, **140**, 3003–3016, doi:10.1175/MWR-D-11-00335.1.

Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, 2002: An improved in situ and satellite SST analysis for climate. *J. Climate*, **15**, 1609–1625, doi:10.1175/1520-0442(2002)015<1609:AIISAS>2.0.CO;2.

Rosner, B., 2000: *Fundamentals of Biostatistics.* Duxbury, 792 pp.

Saha, S., and Coauthors, 2006: The NCEP Climate Forecast System. *J. Climate*, **19**, 3483–3517, doi:10.1175/JCLI3812.1.

——, and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, doi:10.1175/JCLI-D-12-00823.1.