Hindawi Publishing Corporation
Advances in Meteorology
Volume 2013, Article ID 480210, 12 pages
http://dx.doi.org/10.1155/2013/480210

Research Article

Information Content of Seasonal Forecasts in a Changing Climate

Nir Y. Krakauer, Michael D. Grossberg, Irina Gladkova, and Hannah Aizenman

Department of Civil Engineering, The City College of New York, New York, NY 10031, USA

Correspondence should be addressed to Nir Y. Krakauer; nkrakauer@ccny.cuny.edu

Received 14 November 2012; Accepted 19 December 2012

Academic Editor: Klaus Dethloff

We study the potential value to stakeholders of probabilistic long-term forecasts, as quantified by the mean information gain of the forecast compared to climatology. We use as a case study the USA Climate Prediction Center (CPC) forecasts of 3-month temperature and precipitation anomalies made at 0.5-month lead time since 1995. Mean information gain was positive but low (about 2% and 0.5% of the maximum possible for temperature and precipitation forecasts, resp.) and has not increased over time. Information-based skill scores showed similar patterns to other, non-information-based, skill scores commonly used for evaluating seasonal forecasts but tended to be smaller, suggesting that information gain is a particularly stringent measure of forecast quality. We also present a new decomposition of forecast information gain into Confidence, Forecast Miscalibration, and Climatology Miscalibration components. Based on this decomposition, the CPC forecasts for temperature are on average underconfident while the precipitation forecasts are overconfident. We apply a probabilistic trend extrapolation method to provide an improved reference seasonal forecast, compared to the current CPC procedure which uses climatology from a recent 30-year period. We show that combining the CPC forecast with the probabilistic trend extrapolation more than doubles the mean information gain, providing one simple avenue for increasing forecast skill.

1. Introduction

Long-term forecasts offer prospects for enhancing climate readiness and assisting adaptation in sectors including agriculture, fisheries, municipal water supply, hydropower, tourism, and public health [1, 2]. Both statistical and dynamical models have shown some capability for providing long-term forecasts of climate variables such as temperature and precipitation, drawing on "sources of predictability" in the earth system (such as the Southern Oscillation and deep soil moisture) that show persistence or simple dynamics over month to year timescales [3, 4]. Because the skill of long-term forecasts tends to be low, however, they must be accompanied by precise indications of their reliability to be useful to decision makers; believing an overconfident forecast may well lead to worse outcomes than having no forecast available [5–8]. Thus, more so than for synoptic forecasts, useful long-term forecasts cannot be point estimates of a climate quantity but must be presented as a forecast probability distribution [9].

Scoring rules, which provide a metric of skill for a forecast system based on comparing previously issued forecasts to what actually occurred, may be used both to compare different forecast systems and to test improved versions of forecast systems, such as different weightings of ensemble members or methods of bias adjustment [10, 11]. Scoring rules used for point forecasts, such as those based on the squared difference between the forecast and the observed quantity, need modification to be used for probabilistic forecasts. The World Meteorological Organization has recommended using the area under the relative operating characteristics (ROC) curve as a scoring rule for probabilistic forecasts [12]. However, because ROC curves are for dichotomous (yes/no) event outcomes, they do not generalize naturally into a metric for multicategory or continuous climate variables such as temperature and precipitation [4].

Information theory offers simple, general metrics of forecast performance (as information gain (IG) relative to a "no-skill" prior probability distribution). Information gain as a forecast skill score, some of whose advantages are already mentioned by Good [13], has attracted increasing interest in the meteorological community, but has not yet been systematically applied to evaluating and improving long-range forecasts. Roulston and Smith [14] applied a closely related measure they call "ignorance" to evaluating dichotomized short-range temperature forecasts. Bröcker and Smith [15] used IG as the objective function for finding optimal Gaussian kernels to transform point predictions (short- and medium-range ensemble forecasts of temperature) into forecast probability distributions. Benedetti [16] showed that IG is optimal in that it is the unique scoring rule that combines the attributes of being (1) strictly proper, meaning that adopting any forecast instead of the best available one will always decrease the expected score; (2) additive, meaning that it sums across a sequence of forecasts; (3) local, meaning that the probabilities assigned to outcomes that did not occur have no effect on the score. Weijs et al. [17] provided formulas and MATLAB code for a decomposition of the information gain measure into "reliability," "resolution," and "uncertainty" components (along the lines of similar decompositions for other skill scores used in meteorology) and applied this score to evaluating short-term rainfall forecasts for the Netherlands; this decomposition was also independently developed by Tödter [18]. Weijs and van de Giesen [19] extended IG to cases where the observations used for evaluation of the forecast are themselves uncertain. Peirolo [20] showed how to apply the IG score to continuous variables and used it to evaluate short-term ECMWF ensemble predictions of geopotential height and temperature.

While information measures are not new to meteorology applications, they have not been widely applied to long-range forecasts. Acceptance for IG as a metric for scoring and optimizing long-range forecasts therefore requires systematic comparison against other commonly employed measures, such as the correlation coefficient, mean square error, the Brier skill score, and the ranked probability skill score (RPSS).

An additional consideration for scoring seasonal forecasts is what "no-skill" baseline to compare them against. Generally, a climatological mean or probability distribution from some past reference period is used as the baseline. However, in the presence of trends, climatology can give biased estimates of the expected value or probability distribution of the climate variable being forecast [21, 22]. It is therefore of interest to develop baselines that incorporate estimates of the observed trend, building on work done for temperature series by Krakauer [23].

The seasonal forecast product we will evaluate here is the 3-month outlook at 0.5-month lead from the Climate Prediction Center (CPC) of the US National Weather Service (http://www.cpc.ncep.noaa.gov/products/predictions/90day/). On the third Thursday of each month since the end of 1994, CPC releases forecast probabilities of high, low, or near-normal temperature and precipitation over the next 3 months on a 2° grid for the coterminous US (about 200 grid points). The 3 categories of high, low, near-normal are defined as thirds of a climatological distribution based on a recent 30-year period, so that a priori they are said to have equal chances. After each forecast period, CPC also releases verification grids showing the categories observed. We chose to focus on the CPC forecasts because forecasts have been issued in the same format for a relatively long period; archived forecasts and verifications are publicly available, as are (since 2001) discussions of the reasoning behind each forecast; and several previous papers describe the evolution of CPC forecasts over time and present evaluations of their skill [24–27]. The methods presented here should be generally applicable to analyzing the track record of any seasonal forecast product with available verification data.

In this paper, our aims are to (1) estimate the information gain of a seasonal forecast and compare IG to other metrics previously used to evaluate seasonal forecasts, and (2) use trend estimation to better evaluate seasonal forecast skills in a shifting climate and suggest avenues for improving them.

2. Methods

2.1. Information-Based Forecast Skill Metrics. Information metrics for scoring forecast skill are straightforward to interpret and generalize across the type of variable being forecast (e.g., discrete or continuous). If we consider a situation with k possible outcomes indexed $x_1, x_2, \dots, x_k$ (e.g., different temperature quantiles over a given period), the information gained by observing that one of the possible outcomes, x_o, actually took place is given by $-\log(p(x_o))$, where $p(x_i)$ was our prior belief about the probability of outcome x_i. If a forecast shifted our belief about the likelihood of x_o from a reference (climatology) value $p^c(x_o)$ to a value $p^f(x_o)$, we can say that the forecast conveyed information to the extent that the information gained by observing the outcome is less now that we have the forecast. (Thanks to the forecast, the actual outcome became less "surprising"; ignorance is reduced.) The forecast information gain [20] can therefore be defined as

$$\mathrm{IG} = \log\left(p^f(x_o)\right) - \log\left(p^c(x_o)\right) = \log\left(\frac{p^f(x_o)}{p^c(x_o)}\right), \quad (1)$$

which, for a skillful forecast system, will on average be positive.

Alternatively, denote the forecast probability distribution as $\mathbf{f}$, a vector containing the probability of each possible outcome (whose elements f_i, $i = 1, 2, \dots, k$ are therefore nonnegative and sum to 1). Let the verifying probability distribution given by observation be expressed as a vector $\mathbf{o}$, which will be a Kronecker delta if there is no uncertainty in the observation (i.e., the element corresponding to the observed outcome, f_o, will be 1, and all other elements will be 0). Then the relative entropy of the forecast given the observation is defined as

$$\begin{aligned}\mathrm{RE}(\mathbf{o} \,||\, \mathbf{f}) &= \sum_{i=1}^{k} o_i \log \frac{o_i}{f_i} \\ &= \sum_{i=1}^{k} o_i \log o_i - \sum_{i=1}^{k} o_i \log f_i,\end{aligned} \quad (2)$$

where $0 \log 0$ is taken to be zero. In general, RE is a nonnegative function of two probability distributions, also known as

Kullback-Leibler divergence [28], which can be interpreted as the number of bits needed to communicate the observation when its prior probability distribution is given by the forecast. In statistical terms, relative entropy can be understood as a negative log likelihood of the forecast probability distribution given by the observations [29]. The information gain of the forecast over a climatology probability distribution denoted by $\mathbf{c}$ can be written as the reduction in relative entropy afforded by replacing the climatology with the forecast:

$$\begin{aligned} \mathrm{IG} &= \mathrm{RE}(\mathbf{o} \,||\, \mathbf{c}) - \mathrm{RE}(\mathbf{o} \,||\, \mathbf{f}) \\ &= \sum_{i=1}^{k} o_i \log f_i - \sum_{i=1}^{k} o_i \log c_i \\ &= \sum_{i=1}^{k} o_i \log \frac{f_i}{c_i}, \end{aligned} \tag{3}$$

which is the same as (1), generalized to include cases where the observation is uncertain so that $\mathbf{o}$ is not a Kronecker delta.

In practice, assessment of the skill of a forecast system can be based on IG averaged over a large number of forecasts:

$$\langle \mathrm{IG} \rangle = \left\langle \sum_{i=1}^{k} o_i \log \frac{f_i}{c_i} \right\rangle, \tag{4}$$

where $\langle \cdot \rangle$ denotes averaging across some sets of forecasts. $\langle \mathrm{IG} \rangle$ can be thought of as an estimate of the forecast system's expected information gain (with units such as bits or nats, depending on the base of the logarithm taken), or as an average negative log likelihood over the set of forecasts.

To facilitate comparing mean IG to other measures of forecast skill, it may be convenient to normalize it by the maximum possible IG, which would be $\mathrm{RE}(\mathbf{o}||\mathbf{c})$, the IG of a hypothetical perfect forecast that is always identical to the observation $\mathbf{o}$. Two possibilities for computing such an information skill score, which differ with regard to how averaging across forecasts is carried out, are

$$\begin{aligned} \mathrm{ISS}_1 &= \frac{\langle \mathrm{IG} \rangle}{\langle \mathrm{RE}(\mathbf{o} \,||\, \mathbf{c}) \rangle} \\ &= 1 - \frac{\langle \mathrm{RE}(\mathbf{o} \,||\, \mathbf{f}) \rangle}{\langle \mathrm{RE}(\mathbf{o} \,||\, \mathbf{c}) \rangle}, \\ \mathrm{ISS}_2 &= \left\langle \frac{\mathrm{IG}}{\mathrm{RE}(\mathbf{o} \,||\, \mathbf{c})} \right\rangle \\ &= 1 - \left\langle \frac{\mathrm{RE}(\mathbf{o} \,||\, \mathbf{f})}{\mathrm{RE}(\mathbf{o} \,||\, \mathbf{c})} \right\rangle. \end{aligned} \tag{5}$$

ISS_1 and ISS_2 are the same if $\mathrm{RE}(\mathbf{o}||\mathbf{c})$ is constant across forecasts, as would be the case if all elements of $\mathbf{c}$ are identical (an equal-chances reference, which is what we use below). In this paper, we will express skill scores in the ISS_1 form.

Particularly for seasonal forecasts, where because of modest inherent predictability the forecast is often similar to climatology, the following decomposition of information gain may offer insight:

$$\begin{aligned} \mathrm{IG} = {} & \sum_{i=1}^{k} \left(f_i \log f_i - c_i \log c_i \right) + \sum_{i=1}^{k} \left(o_i - f_i \right) \log f_i \\ & - \sum_{i=1}^{k} \left(o_i - c_i \right) \log c_i. \end{aligned} \tag{6}$$

In this new decomposition, the first term (Confidence), which is independent of the outcome, is the difference between the entropy of the reference and forecast distributions; it is high if the issued forecast has much lower entropy (is much more confident) than the reference. The second term (Forecast Miscalibration) should average zero for a well-calibrated forecast; a tendency to positive values suggests an underconfident forecast (since the outcomes forecasted as likely were even more likely to occur than was forecasted), while a tendency to negative values suggests overconfidence (outcomes that were forecasted as likely in fact did not occur as often as expected). The third term (Climatology Miscalibration) is independent of the forecast issued and is zero if the reference distribution is equal chances. This Confidence-Forecast Miscalibration-Climatology Miscalibration decomposition complements the reliability-resolution-uncertainty decomposition of Weijs et al. [17] by highlighting how the amount of information "claimed" by the forecast compares with its actual performance. The Confidence term in the decomposition can be used as the basis for a confidence score (Conf), which is again independent of the observed outcomes:

$$\mathrm{Conf} = 1 - \frac{\left\langle \sum_{i=1}^{k} f_i \log f_i \right\rangle}{\left\langle \sum_{i=1}^{k} c_i \log c_i \right\rangle}. \tag{7}$$

ISS is a *local* score, meaning that only the forecast probability corresponding to the observed outcome has any effect on the score. Locality may be in principle a desirable property, as argued, for example, by Benedetti [16]. If it is nevertheless desired to have forecast probabilities "close" to the outcome to affect the skill score, there is a cumulative variant of ISS, which may be called the ranked information skill score or RISS [18, 30], where, assuming that the event categories are ordered, the observation and forecast vectors $\mathbf{o}, \mathbf{f}$ are replaced by cumulative versions $\mathbf{O}, \mathbf{F}$ according to the following general formula:

$$V_j \equiv \sum_{i=1}^{j} v_i. \tag{8}$$

The formula for RISS is

$$\mathrm{RISS} = 1 - \frac{\langle \mathrm{RE}(\mathbf{O} \,||\, \mathbf{F}) \rangle}{\langle \mathrm{RE}(\mathbf{O} \,||\, \mathbf{C}) \rangle}. \tag{9}$$

Note that strictly speaking RE here is a misnomer, in that in information theory entropies are functions of probability distributions, while the cumulative vectors $\mathbf{O}, \mathbf{F}, \mathbf{C}$ are not probability distributions. In (9), we retain RE to refer to the function defined in (2), with the only formal difference being that the probability distribution vectors are replaced by cumulative versions.

Finally, we give the formulas for alternative metrics for evaluating probabilistic seasonal forecasts that we will compare with IG. The Brier score [31] is given by

$$\mathrm{BS} = \sum_{i=1}^{k} (f_i - o_i)^2. \tag{10}$$

This can be considered as a second-order polynomial approximation to $\mathrm{RE}(\mathbf{o} \| \mathbf{f})$ [16, 17]. We average across a set of forecasts and normalize by the reference forecast's score to produce a Brier skill score analogous to ISS_1 as follows:

$$\mathrm{BSS} = 1 - \frac{\left\langle \sum_{i=1}^{k} (f_i - o_i)^2 \right\rangle}{\left\langle \sum_{i=1}^{k} (c_i - o_i)^2 \right\rangle}. \tag{11}$$

The cumulative variant of BS is called the ranked probability score [32], which as with RISS replaces $\mathbf{o}, \mathbf{f}$ by cumulative versions $\mathbf{O}, \mathbf{F}$ as follows:

$$\mathrm{RPSS} = 1 - \frac{\left\langle \sum_{i=1}^{k} (F_i - O_i)^2 \right\rangle}{\left\langle \sum_{i=1}^{k} (C_i - O_i)^2 \right\rangle}. \tag{12}$$

RPSS may also be defined with other positive exponents replacing 2 in (12) [33].

The Heidke score [34] was formulated for deterministic forecasts and is used for probabilistic forecasts by replacing these with deterministic forecasts for the most probable outcome. The Heidke score HS for a forecast may be defined to be 1 if the outcome predicted by the deterministic forecast takes place and $-1/(k-1)$ if it does not (where k is the number of possible outcomes). If the probabilistic forecast has no most probable category (e.g., an equal-chances forecast), HS is taken to be 0. Given HS, the corresponding normalized skill score would be

$$\mathrm{HSS} = \frac{\langle \mathrm{HS}(\mathbf{f}, \mathbf{o}) - \mathrm{HS}(\mathbf{c}, \mathbf{o}) \rangle}{\langle 1 - \mathrm{HS}(\mathbf{c}, \mathbf{o}) \rangle}. \tag{13}$$

Clearly, all nuance conveyed by the confidence of a probabilistic forecast is lost in HSS; a forecast vector $\mathbf{f}$ of $(0.9, 0.05, 0.05)$, for example, would always get the same Heidke score as one of $(0.4, 0.3, 0.3)$.

2.2. Forecast and Verification Data. On the third Thursday of each month since the end of 1994, CPC releases forecast probabilities of high, low, or near-normal temperature and precipitation over the next 3 months on a 2° grid for the coterminous US (232 grid points); for example, January-February-March mean temperature and precipitation are forecast in mid-December. The 3 categories of high, low, and near normal are defined as thirds of a climatological distribution based on a recent 30-year period, so that a priori they are said to have equal chances; this was taken as our reference probability distribution $\mathbf{c}$. Verification observations are also taken from the CPC and are in the same categories and grid as the forecast; we neglect any error in these observations. Forecast-verification sets were available for 209 consecutive start months, from January 1995 to May 2012.

2.3. A Trend-Following Baseline. Given significant recent trends in climate quantities that have led to substantial shifts in probability distributions compared to past climatology, we considered improving on the equal-chances reference forecast vector by updating the climatological probability distribution each year since 1995 based on observations as

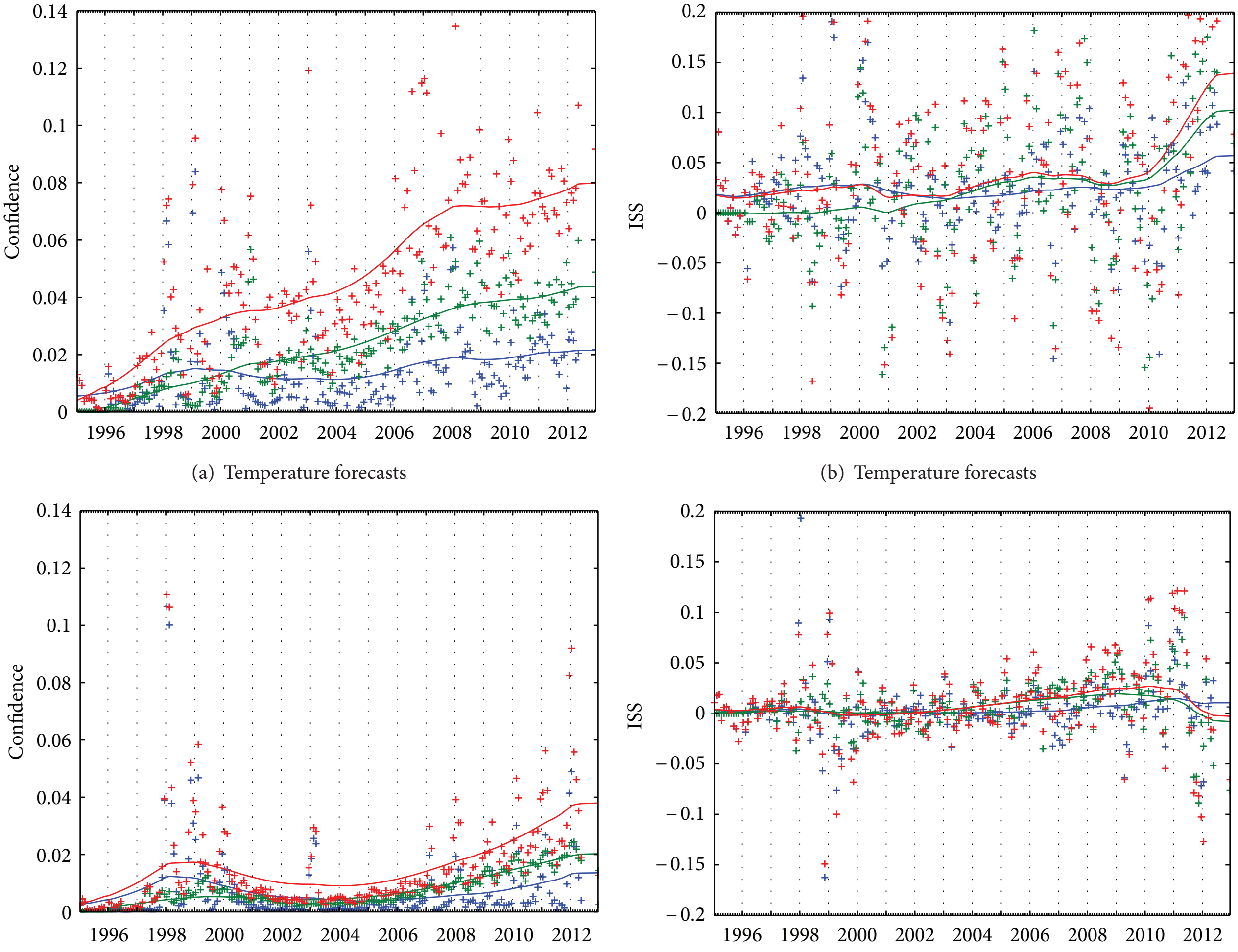
$$\mathbf{f}^{\mathrm{trend}}(t+1) = (1-\alpha)\,\mathbf{f}^{\mathrm{trend}}(t) + \alpha \mathbf{o}(t). \tag{14}$$

Thus, the Trend forecast $\mathbf{f}^{\mathrm{trend}}$ evolves with time to follow observations. The parameter α sets the weight given to an individual observation in the forecast for next year, and its optimal value is related to the ratio between the magnitude of the climate trend to the magnitude of the year-to-year variability [35]. In practice, we determine a value for α for each year and separately for temperature and precipitation (but uniform across grid points and seasons) based on maximizing the log likelihood (i.e., minimizing the RE) for prediction over previous periods within the range $0.02 < \alpha < 0.08$. This corresponds to a characteristic averaging timescale $1/\alpha$ of 12.5 to 50 years, similar to that typically used for the construction of optimal climate normals [36]. The initial distribution $\mathbf{f}^{\mathrm{trend}}$ $(t = 1995)$ was set to equal chances, and for the first year, a default value of $\alpha = 0.04$ was used. We found that after several years, the optimized α stabilized at about 0.06 for temperature and 0.03 for precipitation.

To explore whether CPC forecasts $\mathbf{f}^{\mathrm{CPC}}$ could benefit from incorporating information from $\mathbf{f}^{\mathrm{trend}}$, we constructed a simple Combined forecast via the naive Bayesian approach as follows:

$$f_i^{\mathrm{comb}} = \frac{f_i^{\mathrm{CPC}} f_i^{\mathrm{trend}}}{\sum_{j=1}^{k} f_j^{\mathrm{CPC}} f_j^{\mathrm{trend}}}. \tag{15}$$

2.4. Statistical Inference. Any measure of forecast skill is expected to vary from event to event. In order to declare a forecast system as having positive average skill, or one forecast system as having more skill than another, it is necessary to estimate the uncertainty of the average skill, with the set of available forecast-observation pairs viewed as samples from the stochastic forecast and climate systems [37]. This task is complicated by the expected temporal correlation of forecast skill, since forecast periods overlap and since climate events that affect the observed values, such as El Niño episodes, persist for multiple months [38]. We estimated the uncertainty in the mean skill of a forecast as the standard error of the mean term in a fit of a low-order seasonal autoregressive moving average (SARMA) process to the time series. To determine whether there was a significant monotonic (not necessarily linear) trend in forecast skill over the study period, we used the nonparametric Mann-Kendall test on the SARMA residuals. $P < 0.05$ (two-tailed) was set as the threshold for significance.



(a) Temperature forecasts

(b) Temperature forecasts

(c) Precipitation forecasts

(d) Precipitation forecasts

Figure 1: (a) Mean confidence score for the temperature forecasts (markers; lines are smoothed based on local linear regression with a bandwidth of 24 months). (b) Mean information skill score for the temperature forecasts, (c)-(d) same as (a)-(b), but for precipitation forecasts. The color scheme is blue for the CPC forecast, green for the Trend forecast, and red for the Combined forecast.

Table 1: Average confidence and information gain for forecasts and trend extrapolation.

	Temperature			Precipitation		
	CPC	Trend	Comb	CPC	Trend	Comb
1995–2012						
Conf	0.0144	0.0227	0.0463	0.0068	0.0072	0.0150
ISS	0.0236	0.0215	0.0332	0.0031	0.0071	0.0090
2003–2012						
Conf	0.0164	0.0326	0.0627	0.0060	0.0094	0.0170
ISS	0.0241	0.0402	0.0461	0.0047	0.0130	0.0159

3. Results

3.1. Evolution of Forecast Confidence and Skill. Figure 1 shows the confidence score Conf and the information skill score ISS$_1$ averaged over all grid points for each month. The confidence of the trend-based probability distribution is zero for the first year and gradually increases as more years of history become available to allow estimation of trends, generally surpassing the CPC forecast over recent years. The CPC forecasts' confidence is more variable, peaking on occasions such as strong El Niño episodes when the forecasters believed that seasonal climate had more predictability. The combined forecast, as expected from its construction, consistently has more confidence than either of its components.

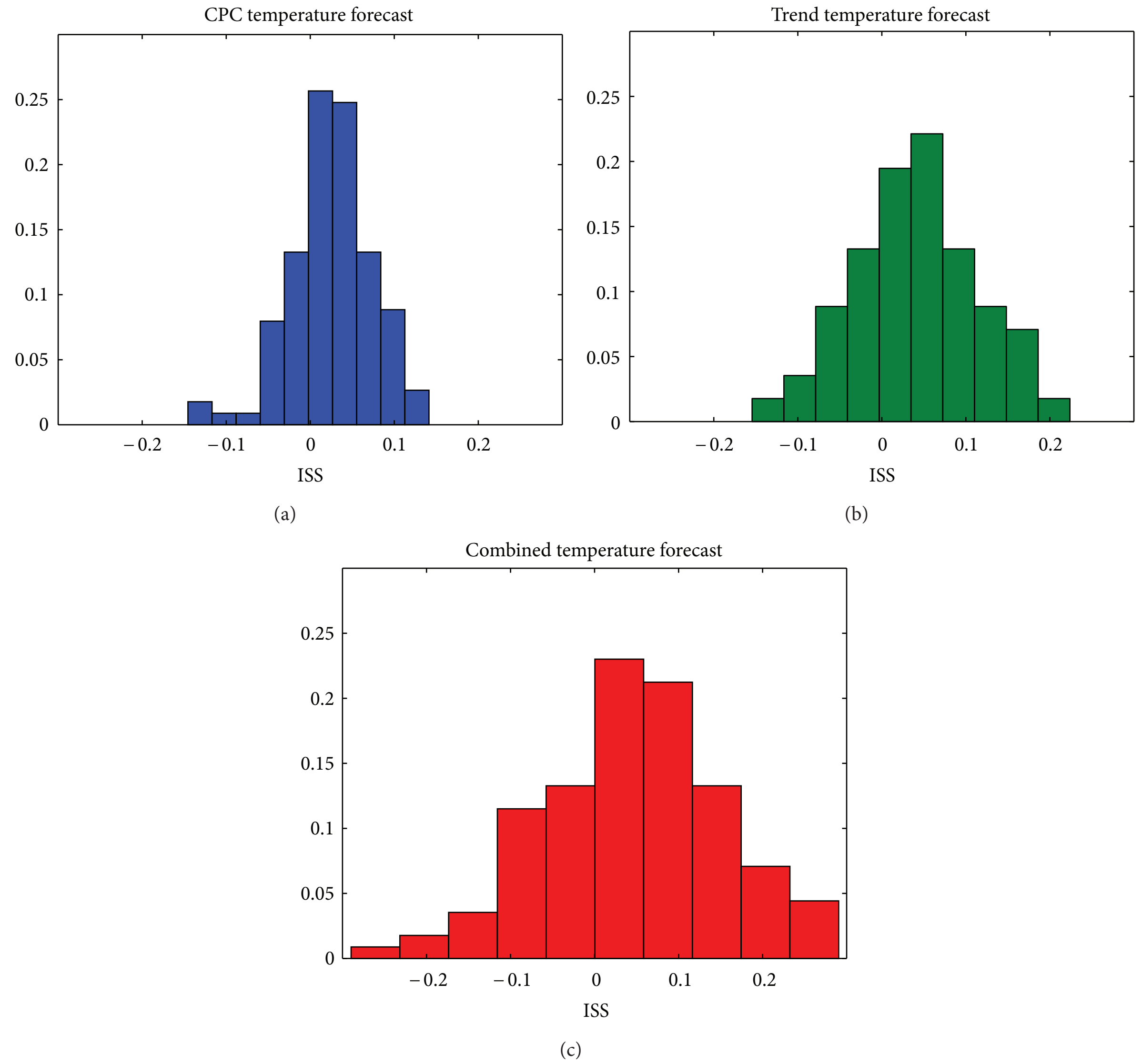
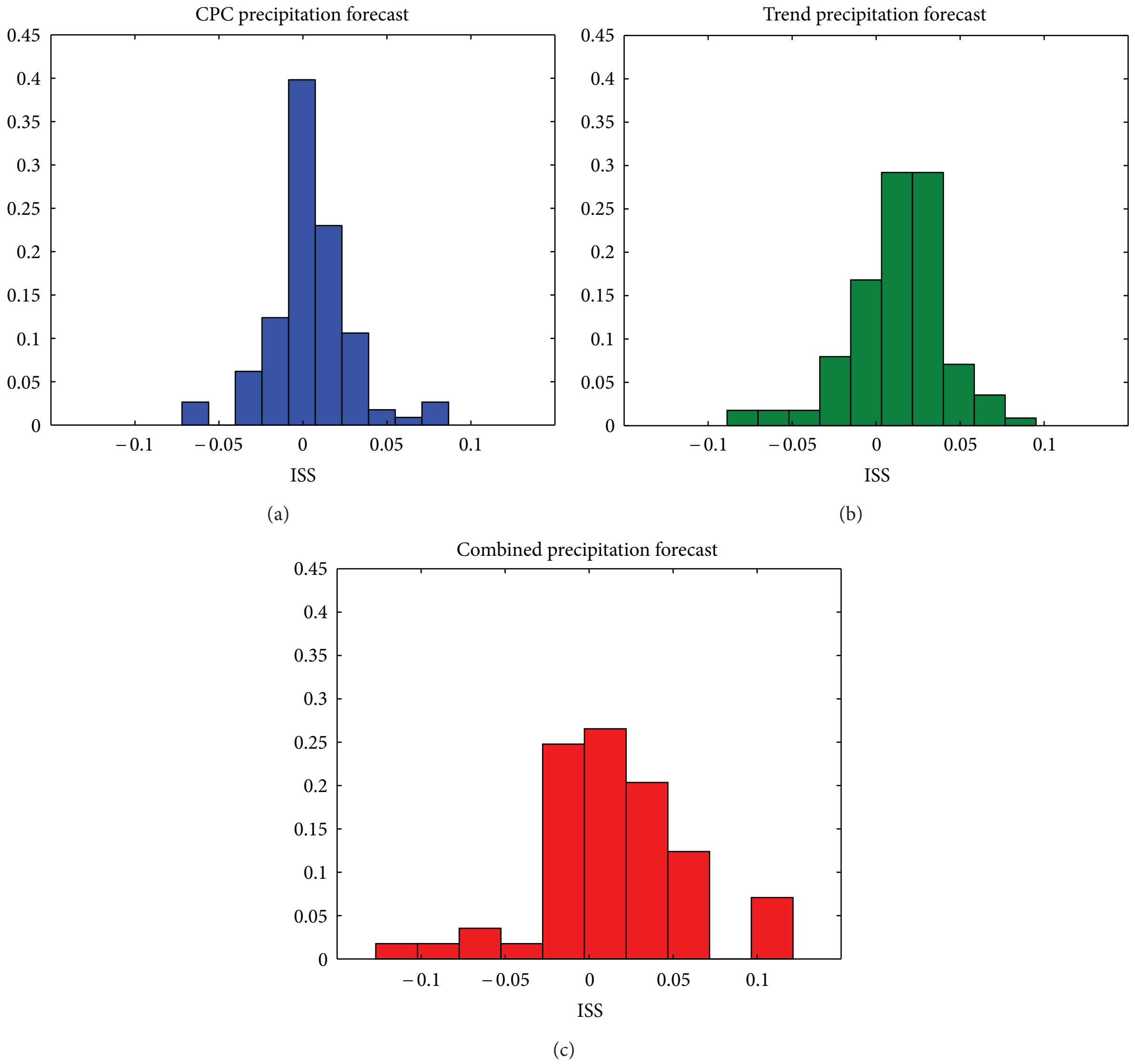


Figure 2: Histogram of mean monthly ISS for temperature forecasts: (a) CPC; (b) Trend; (c) Combined.

ISS is more variable than Conf since it depends on observations as well as on the forecast system; for many months, ISS is negative, meaning that the forecasts performed worse than an equal-chances prediction (Figures 1(b) and 1(d)). Overall, the CPC temperature forecasts have positive ISS for 73% of months since 2003, and the precipitation forecasts have positive ISS only for 58% of months; the corresponding figures for the Trend forecast are 72% and 73% (Figures 2 and 3). For temperature, the warming picked up by the Trend forecast gives it generally more skill than the CPC forecast since around 2004, with a dip in 2009 when a cool period led trend to be a poor basis for forecasting. For precipitation, skill as well as confidence is lower than that for temperature, since both trends and the persistent factors considered by the CPC forecasters are less strong indicators. Here also Trend on average outperforms the CPC forecast since 2004, with a noteworthy exception since late 2011 when dry conditions have broken with a trend toward increasing precipitation in the north-central USA.

Overall average values for Conf and ISS are given in Table 1. Since the Trend has built up confidence over time, averages since 2003 as well as for the entire CPC forecast period (since 1995) are given. For temperature, the CPC forecast has higher ISS than Conf suggesting that the CPC forecasters have tended to be underconfident. The Trend forecast has almost the same mean ISS for the entire period and higher ISS since 2003, and seems to be fairly well calibrated (ISS is similar to Conf), suggesting that the RE-based optimization of the parameter α has been successful in choosing appropriate values. For precipitation, the CPC forecast has less confidence than for temperature, but ISS is even lower, suggesting overconfidence for the CPC precipitation forecasts. Trend also gave less confidence for precipitation than for temperature, but higher confidence and ISS than



(a)

(b)

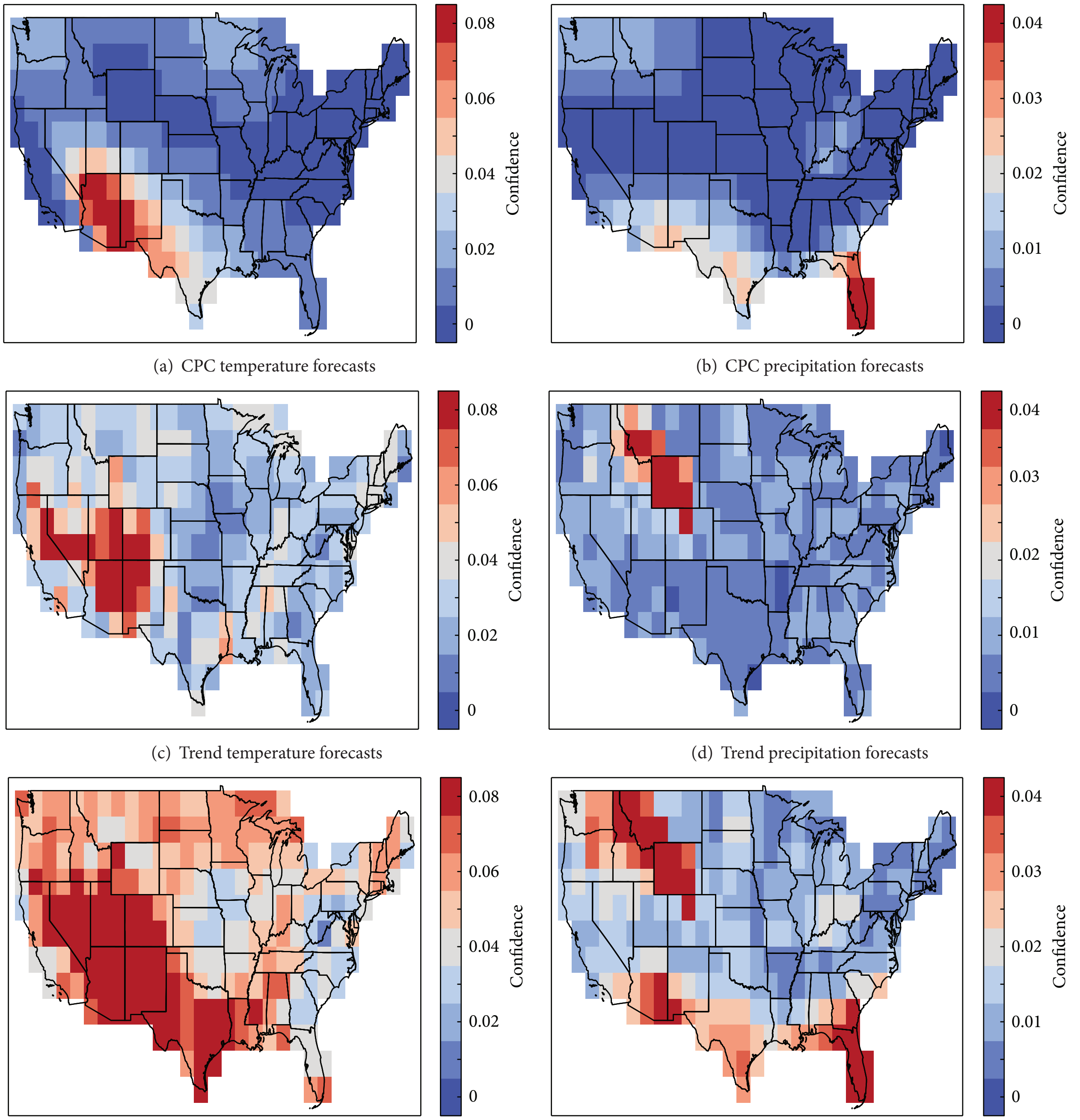
(c)

Figure 3: Same as Figure 2, but for precipitation forecasts.

that for CPC. The Combined forecast ISS was on average better than either CPC or Trend, suggesting that the two contain some independent information, but less than the combined forecast's confidence. This overconfidence for the combined forecast is expected since naive Bayesian combination assumes that the different components (CPC and Trend) are entirely uncorrelated; in fact, the CPC forecasters do make some use of trends, so the two forecasts are not independent and their optimal combination should have lower confidence to reflect this. The Combined forecast also had a somewhat lower percentage of months with positive ISS than the Trend forecast—69% for temperature and 62% for precipitation—again plausibly reflecting its overconfidence.

3.2. Spatial Patterns in Confidence and Skill. It is of interest to see what regions have accounted for the CPC forecast and Trend confidence and skill. Figures 4 and 5 show mean Conf and ISS for the two probability distributions by grid cell, averaged since 2003. The CPC temperature forecasts show the highest confidence in an area stretching from the Great Basin to the Texas coast, largely overlapping with where the Trend confidence is the highest, although the Trend shows some confidence and skill for parts of the east. For precipitation, the CPC forecast confidence is concentrated along the southern margin of the US from Arizona to Florida, where winter warm, dry conditions are associated with La Niña and cool, wet conditions with El Niño. The Trend forecast focuses on the northern Rockies and Great Plains, where there has been wetting.

As another depiction of the geographic variation in forecast skill, Figure 6 shows the mean ISS (since 2003) for quarters of the coterminous USA, split at 39°N and 99°W. For temperature, the CPC forecasts have the most skill for the southeast and southwest, and combining then with the Trend yields improved forecasts for all quarters. For precipitation, the CPC forecasts on average only have skill in



(a) CPC temperature forecasts

(b) CPC precipitation forecasts

(c) Trend temperature forecasts

(d) Trend precipitation forecasts

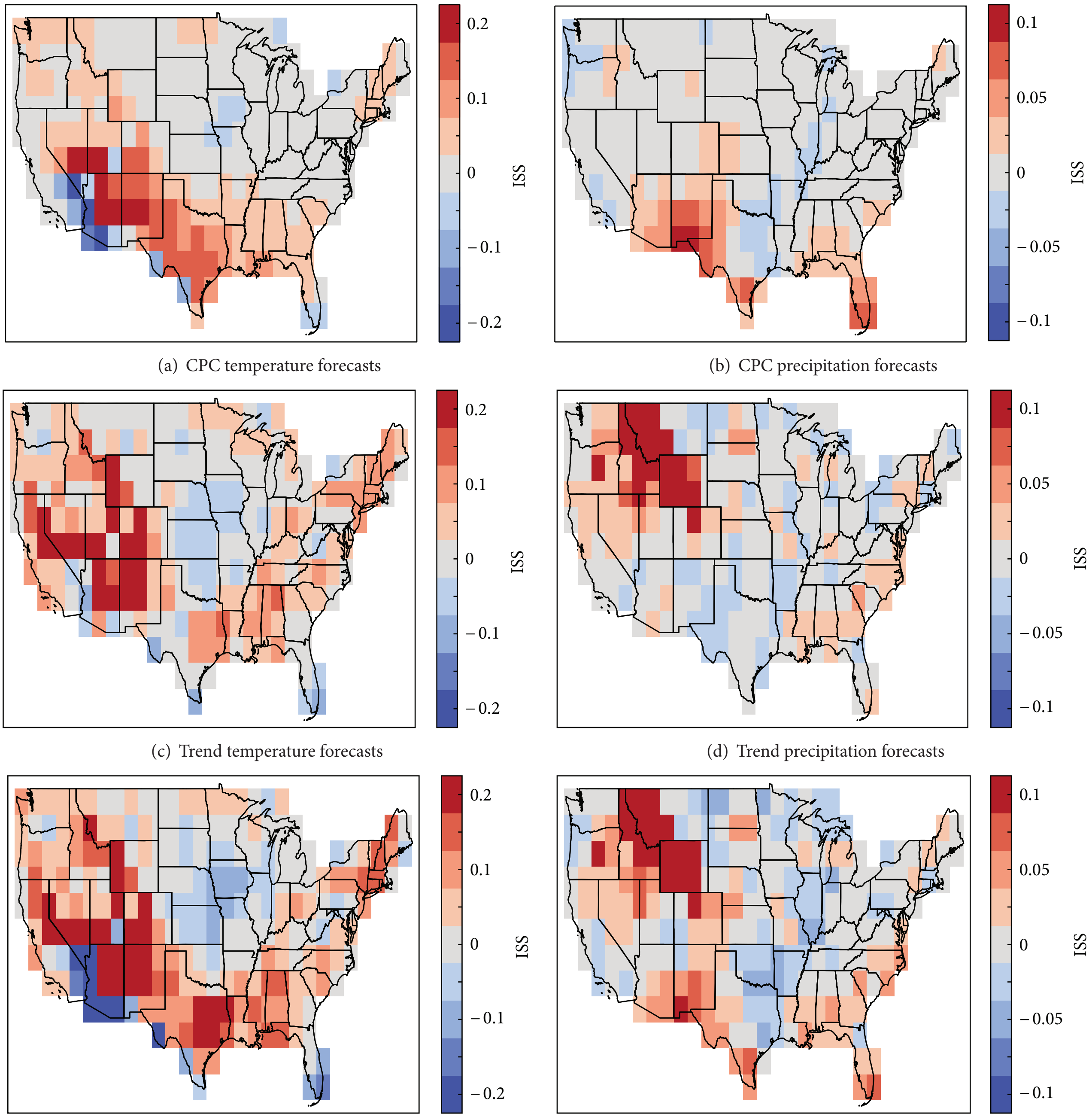
(e) Combined temperature forecasts

(f) Combined precipitation forecasts

FIGURE 4: Mean confidence score for the CPC forecasts of (a) temperature and (b) precipitation. (c)-(d) Same, but for Trend forecast. (e)-(f) Same, but for Combined forecast.

the southwest whereas the Trend has skill in the northwest, and the Combined forecast has both these areas of strength.

3.3. Comparison across Skill Metrics. Table 2 shows how the CPC forecast compares with the Trend and Combined forecasts as judged by the different skill metrics introduced in Section 2 (all averaged since 2003). ISS is generally the most demanding skill score (with lower values than the other metrics). BSS is very similar to but slightly higher than ISS, which is consistent with the Brier score being a second-order approximation to IG that penalizes wrong predictions less. The ranked versions of ISS and BSS mostly give higher skill scores, which makes sense since the ranking gives credit for being "closer" to the observed outcome even if the observed outcome was not forecasted as being probable. HSS has the highest skill score of all the metrics considered. All skill scores agreed in finding the Trend more skillful than the CPC forecast for both temperature and precipitation, and the combined forecast somewhat more skillful still.



(a) CPC temperature forecasts

(b) CPC precipitation forecasts

(c) Trend temperature forecasts

(d) Trend precipitation forecasts

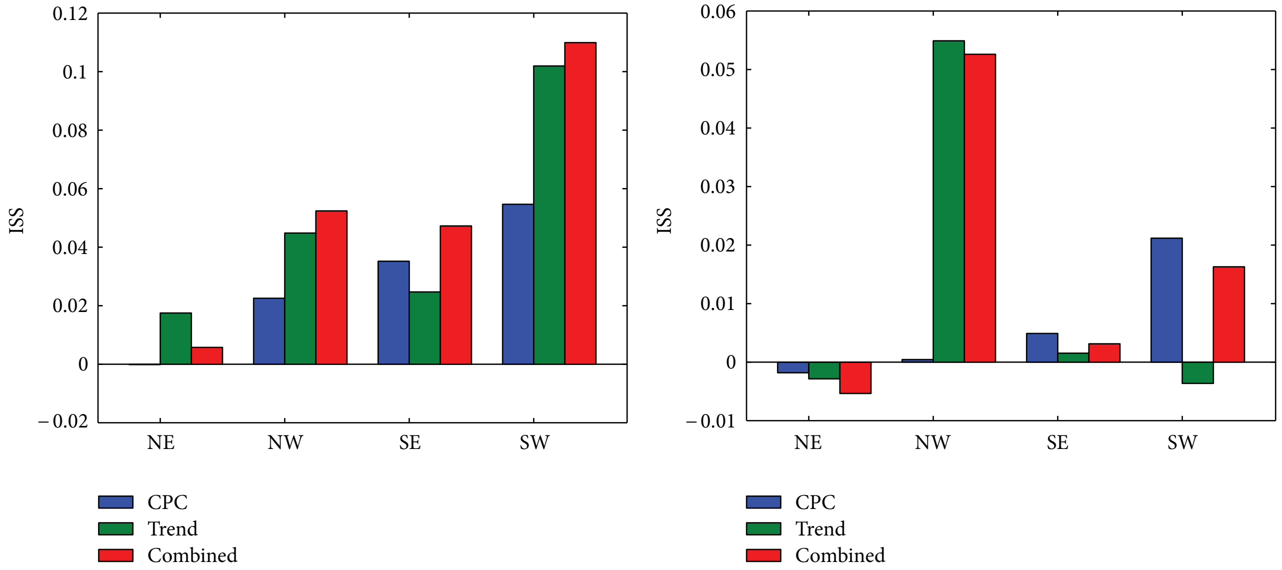
(e) Combined temperature forecasts

(f) Combined precipitation forecasts

Figure 5: Mean information skill score for the CPC forecasts of (a) temperature and (b) precipitation. (c)-(d) Same, but for Trend forecast. (e)-(f) Same, but for Combined forecast.

Table 2: Skill scores for forecasts and trend extrapolation.

	Temperature			Precipitation		
	CPC	Trend	Comb	CPC	Trend	Comb
ISS	0.0241	0.0402	0.0461	0.0047	0.0130	0.0159
RISS	0.1341	0.2518	0.3506	0.0012	0.0523	0.0505
BSS	0.0274	0.0458	0.0529	0.0053	0.0152	0.0185
RPSS	0.0407	0.0725	0.0825	0.0080	0.0226	0.0276
HSS	0.2163	0.3705	0.3714	0.0895	0.2980	0.3071



(a) Temperature

(b) Precipitation

Figure 6: Mean information skill score by quarter of the USA for forecasts of (a) temperature and (b) precipitation.

Fitting seasonal autoregressive models to the skill score time series showed that there was no significant trend in the CPC forecast skill either since 1995 or since 2003, regardless of the metric chosen (not shown). The mean CPC forecast skill for precipitation was not significantly different from zero under all metrics except HSS while the mean Trend skill was greater than zero for the period since 2003 under all metrics; for temperature, both CPC and Trend had significant skill under all metrics (not shown).

4. Discussion

4.1. Seasonal Forecast Skill: Comparison with Previous Assessments. Many of our results—for example, that skill for precipitation is lower than that for temperature and that CPC does not optimally account for trends—are largely consistent with previous assessments of the CPC seasonal forecasts [24–27], albeit these did not conduct a quantitative comparison with a probabilistic trend extrapolation. (Peng et al. [27] compared CPC with a deterministic trend forecast based on the most common category seen over the previous few years, but such a deterministic forecast cannot be used for comparing probabilistic skill scores.) Underestimation of trends also appears to be characteristic of some other operational seasonal forecasts [39]. The lack of a significant improvement in the CPC forecasts over time is discouraging, particularly since over the last few years the CPC forecast methodology has been revised to more objectively weight the different potential sources of predictability [25]. We show that the CPC forecast is indeed quantitatively dominated by a trend extrapolation (our Trend forecast), and that adjusting the CPC forecast to include the trend, even in a simplistic way, results in substantially improved average skill relative to an equal-chances baseline.

If there is specific reason to believe either that no trend in the variable being forecast exists or that trends observed over recent years have now reversed, then this information should be incorporated into the reference forecast instead of relying on trend extrapolation blindly. For example, while temperature increased rather linearly since the 1970s [21], precipitation may not be well characterized by a consistent trend, as hinted by the poor performance of our trend extrapolation for precipitation during the 2011-2012 drought. Further research on the performance of trend extrapolation for different climate variables, regions, and time periods is warranted.

Once seasonal forecasts do incorporate trends appropriately, the time-varying trend extrapolation may in fact be a more appropriate reference probability distribution $\mathbf{c}$ than equal chances based on climatology, since the expectation is that a seasonal forecasting system should be able to use knowledge of specific current conditions to outperform mere trend extrapolation. The confidence score we introduced, based on our decomposition of the information gain into Confidence, Forecast Miscalibration, and Climatology Miscalibration components, could be used in conjunction with ISS to help calibrate probabilistic forecasts such as CPC's.

4.2. Information Skill Scores for Assessing Seasonal Forecasts. The comparisons shown here suggest that there is great, relatively systematic variation in the skill score generated by different metrics, even when normalized to a common scale (where 0 corresponds to no skill and 1 to a perfect forecast). ISS is generally the more stringent skill score; for example, HSS averages more than a factor of 10 greater than ISS for the CPC seasonal forecasts. This suggests the need for more exploration of how well the different skill scores correspond to user requirements; in general, no single skill score can be expected to capture all aspects of forecast performance, which can only be completely described by the full joint probability distribution of forecasts and observations [40]. Information measures of forecast skill have the advantage of a clear theoretical basis in terms of the underlying joint probability distribution [41] and should be added to forecast verification software tools such as the Ensemble Verification System [42] in order to facilitate comparing them with currently used metrics.

5. Conclusions

Information gain measures show that at least the CPC seasonal temperature forecast has measurable skill, but that for it and the precipitation forecast the skill can be at least doubled by adjusting the probability distribution based on recent trends. Comparing seasonal forecasts to probabilistic trend extrapolation and comparing confidence scores to information gain (where the two should on average be equal for a well-calibrated forecast) are tools introduced here that should help improve seasonal forecasts substantially.

Acknowledgments

The authors gratefully acknowledge support from NOAA under Grants NA11SEC4810004 and NA12OAR4310084. All statements made are the views of the authors and not the opinions of the funding agency or the US government.

References

[1] E. Klopper, C. H. Vogel, and W. A. Landman, “Seasonal climate forecasts-potential agricultural-risk management tools?” *Climatic Change*, vol. 76, no. 1-2, pp. 73–90, 2006.

[2] A. Troccoli, “Seasonal climate forecasting,” *Meteorological Applications*, vol. 17, no. 3, pp. 251–268, 2010.

[3] J. Namias, “Remarks on the potential for long-range forecasting,” *Bulletin of the American Meteorological Society*, vol. 66, no. 2, pp. 165–173, 1985.

[4] National Research Council, “Assessment of intraseasonal to interannual climate prediction and predictability,” Technical Report, National Research Council, Washington, DC, USA, 2010.

[5] T. C. Pagano, H. C. Hartmann, and S. Sorooshian, “Using climate forecasts for water management: Arizona and the 1997-1998 El Niño,” *Journal of the American Water Resources Association*, vol. 37, no. 5, pp. 1139–1153, 2001.

[6] K. Wernstedt and R. Hersh, “Climate forecasts in flood planning: promise and ambiguity,” *Journal of the American Water Resources Association*, vol. 38, no. 6, pp. 1703–1713, 2002.

[7] S. Apipattanavis, F. Bert, G. Podestá, and B. Rajagopalan, “Linking weather generators and crop models for assessment of climate forecast outcomes,” *Agricultural and Forest Meteorology*, vol. 150, no. 2, pp. 166–174, 2010.

[8] P. Block, “Tailoring seasonal climate forecasts for hydropower operations,” *Hydrology and Earth System Sciences*, vol. 15, no. 4, pp. 1355–1368, 2011.

[9] A. H. Murphy and R. L. Winkler, “Probability forecasting in meterology,” *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 489–500, 1984.

[10] R. L. Winkler, J. Muñoz, J. L. Cervera et al., “Scoring rules and the evaluation of probabilities,” *Test*, vol. 5, no. 1, pp. 1–60, 1996.

[11] T. Gneiting and A. E. Raftery, “Strictly proper scoring rules, prediction, and estimation,” *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

[12] World Meteorological Organization, “Standardised Verification System (SVS) for Long-Range Forecasts (LRF),” Technical Report. New Attachment II-9 to the Manual on the GDPS WMO-No. 485, World Meteorological Organization, Geneva, Switzerland, 2002, volume 1.

[13] I. J. Good, “Rational decisions,” *Journal of the Royal Statistical Society B*, vol. 14, no. 1, pp. 107–114, 1952.

[14] M. S. Roulston and L. A. Smith, “Evaluating probabilistic forecasts using information theory,” *Monthly Weather Review*, vol. 130, no. 6, pp. 1653–1660, 2002.

[15] J. Bröcker and L. A. Smith, “From ensemble forecasts to predictive distribution functions,” *Tellus A*, vol. 60, no. 4, pp. 663–678, 2008.

[16] R. Benedetti, “Scoring rules for forecast verification,” *Monthly Weather Review*, vol. 138, no. 1, pp. 203–211, 2010.

[17] S. V. Weijs, R. van Nooijen, and N. van de Giesen, “Kullback-Leibler divergence as a forecast skill score with classic reliability-resolution-uncertainty decomposition,” *Monthly Weather Review*, vol. 138, pp. 3387–3399, 2010.

[18] J. Tödter, *New aspects of information theory in probabilistic forecast verification [M.S. thesis]*, Goethe University, Frankfurt, Germany, 2011.

[19] S. V. Weijs and N. van de Giesen, “Accounting for observational uncertainty in forecast verification: an information-theoretical view on forecasts, observations, and truth,” *Monthly Weather Review*, vol. 139, no. 7, pp. 2156–2162, 2011.

[20] R. Peirolo, “Information gain as a score for probabilistic forecasts,” *Meteorological Applications*, vol. 18, no. 1, pp. 9–17, 2011.

[21] R. E. Livezey, K. Y. Vinnikov, M. M. Timofeyeva, R. Tinker, and H. M. van den Dool, “Estimation and extrapolation of climate normals and climatic trends,” *Journal of Applied Meteorology and Climatology*, vol. 46, no. 11, pp. 1759–1776, 2007.

[22] P. C. D. Milly, J. Betancourt, M. Falkenmark et al., “Climate change: stationarity is dead: whither water management?” *Science*, vol. 319, no. 5863, pp. 573–574, 2008.

[23] N. Y. Krakauer, “Estimating climate trends: application to United States plant hardiness zones,” *Advances in Meteorology*, vol. 2012, Article ID 404876, 9 pages, 2012.

[24] D. S. Wilks, “Diagnostic verification of the Climate Prediction Center long-lead outlooks, 1995–9,” *Journal of Climate*, vol. 13, no. 13, pp. 2389–2403, 2000.

[25] E. A. O’Lenic, D. A. Unger, M. S. Halpert, and K. S. Pelman, “Developments in operational long-range climate prediction at CPC,” *Weather and Forecasting*, vol. 23, no. 3, pp. 496–515, 2008.

[26] R. E. Livezey and M. M. Timofeyeva, “The first decade of long-lead U.S. seasonal forecasts,” *Bulletin of the American Meteorological Society*, vol. 89, no. 6, pp. 843–854, 2008.

[27] P. Peng, A. Kumar, M. S. Halpert, and A. G. Barnston, “An analysis of CPC’s operational 0.5-month lead seasonal outlooks,” *Weather and Forecasting*, vol. 27, no. 4, pp. 898–917, 2012.

[28] S. Kullback and R. A. Leibler, “On information and sufficiency,” *Annuals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[29] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, NY, USA, 1991.

[30] B. Ahrens and A. Walser, "Information-based skill scores for probabilistic forecasts," *Monthly Weather Review*, vol. 136, no. 1, pp. 352–363, 2008.

[31] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, pp. 1–3, 1950.

[32] A. H. Murphy, "A note on the ranked probability score," *Journal of Applied Meteorology*, vol. 10, no. 1, pp. 155–156, 1971.

[33] W. A. Müller, C. Appenzeller, F. J. Doblas-Reyes, and M. A. Liniger, "A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes," *Journal of Climate*, vol. 18, no. 10, pp. 1513–1523, 2005.

[34] P. Heidke, "Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst," *Geografiska Annaler*, vol. 8, pp. 301–349, 1926.

[35] V. Naumov and O. Martikainen, "Exponentially weighted simultaneous estimation of several quantiles," *World Academy of Science, Engineering and Technology*, vol. 8, pp. 563–568, 2007.

[36] J. Huang, H. M. van den Dool, and A. G. Barnston, "Long-lead seasonal temperature prediction using optimal climate normals," *Journal of Climate*, vol. 9, no. 4, pp. 809–817, 1996.

[37] A. A. Bradley, S. S. Schwartz, and T. Hashino, "Sampling uncertainty and confidence intervals for the Brier score and Brier skill score," *Weather and Forecasting*, vol. 23, no. 5, pp. 992–1006, 2008.

[38] D. S. Wilks, "Sampling distributions of the Brier score and Brier skill score under serial dependence," *Quarterly Journal of the Royal Meteorological Society*, vol. 136, no. 653, pp. 2109–2118, 2010.

[39] A. G. Barnston, S. Li, S. J. Mason, D. G. Dewitt, L. Goddard, and X. Gong, "Verification of the first 11 years of IRI's seasonal climate forecasts," *Journal of Applied Meteorology and Climatology*, vol. 49, no. 3, pp. 493–520, 2010.

[40] A. H. Murphy, "Forecast verification: its complexity and dimensionality," *Monthly Weather Review*, vol. 119, no. 7, pp. 1590–1601, 1991.

[41] S. V. Weijs, G. Schoups, and N. Van De Giesen, "Why hydrological predictions should be evaluated using information theory," *Hydrology and Earth System Sciences*, vol. 14, no. 12, pp. 2545–2558, 2010.

[42] J. D. Brown, J. Demargne, D. J. Seo, and Y. Liu, "The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations," *Environmental Modelling and Software*, vol. 25, no. 7, pp. 854–872, 2010.