# Estimation of seasonal precipitation tercile-based categorical probabilities from ensembles

MICHAEL K. TIPPETT, ANTHONY G. BARNSTON AND ANDREW W. ROBERTSON

*International Research Institute for Climate and Society, Palisades, NY, USA*

April 27, 2006

ABSTRACT

Counting the number of ensemble members in each tercile-based category is a simple non-parametric method of using a forecast ensemble to assign categorical probabilities. Parametric methods include estimating categorical probabilities from fitted distributions and from generalized linear regression models. Here we investigate the effect of sampling error due to finite ensemble size on the accuracy of counting and parametric methods, focusing on tercile-based categories. The methods are first compared in an idealized setting where analytical results show the dependence of sampling error on method, ensemble size, and level of predictability. We find that the analytical results provide a good description of the behavior of seasonal precipitation probabilities simulated by a general circulation model with parametric methods being generally more accurate than the counting method. We also show how the accuracy of the categorical probability estimates affects the rank probability skill score. In addition to determining the relative accuracies of the different methods, the analysis quantifies the relative importance of the ensemble mean and variance in determining tercile probabilities, with ensemble variance being shown to be a weak factor in determining seasonal precipitation probabilities.

## 1. Introduction

Seasonal climate forecasts are necessarily probabilistic, and forecast information is most completely characterized by a probability density function (pdf). Estimation of the forecast pdf is required to measure predictability and to issue accurate forecasts. For reliable forecasts, the difference between the climatological and forecast pdfs represents predictability, and several measures of this difference have been developed to quantify predictability (Kleeman 2002; DelSole 2004; Tippett et al. 2004; DelSole and Tippett 2006). Quantile probabilities provide a coarse-grained description of the forecast and climatological pdfs that is appropriate for ensemble methods with relatively small sample size. The International Research Institute for Climate and Society (IRI) issues seasonal forecasts of precipitation and temperature in the form of tercile-based categorical probabilities (hereafter called tercile probabilities)–that is, the probability of the below-normal, normal and above-normal categories (Barnston et al. 2003). Forecasts that differ from equal-odds probabilities, to the extent that they are reliable, are indications of predictability in the climate system. Accurate estimation of quantile probabilities is important both for quantifying seasonal predictability and for making climate forecasts.

In single-tier seasonal climate forecasts, initial conditions of the ocean-land-atmosphere system are the source of predictability, and ensembles of coupled model forecasts provide samples of the model atmosphere-land-ocean system consistent with the initial condition, its uncertainty and the internal variability of the coupled model. In simulations and two-tier seasonal forecasts, an ensemble of atmospheric general circulation models (GCMs) provides a sample of equally likely model atmospheric responses to a particular configuration of sea surface temperature (SST). Tercile probabilities must be estimated from finite ensembles in either system. A simple nonparametric estimate of the tercile probabilities is the fraction of ensemble members in each category. Alternatively, the entire forecast pdf including tercile probabilities can be estimated by modeling the

ensemble as a sample from a pdf with some adjustable parameters–here, a Gaussian distribution. The counting method has the advantage of making no assumptions about the forecast pdf. Both approaches are affected by sampling error due to finite ensemble size, though to different degrees. This paper is about the impact of sampling error on parametric and nonparametric estimates of simulated and forecast tercile probabilities for seasonal precipitation totals. We analyze precipitation because of its societal importance and because, even on seasonal time-scales, its distribution is farther from being Gaussian and hence more challenging to describe than other quantities like temperature and geopotential height which have been previously examined.

In this paper we present analytical descriptions of the accuracy of the counting and Gaussian tercile probability estimators. These analytical results facilitate the comparison of the counting and Gaussian estimates and show how the accuracy of the estimators increases as ensemble size and predictability level increase. The analytical results support previous empirical results showing the advantage of the parametric estimators. Wilks (2002) found that modeling numerical weather prediction ensembles with Gaussian or Gaussian mixture distributions gave more accurate estimations of quantile values than counting, especially for quantiles near the extremes of the distribution. Kharin and Zwiers (2003) used Monte Carlo simulations to show that a Gaussian fit estimate was more accurate than counting for Gaussian distributed forecast variables.

We show how the accuracy of the tercile probability estimates affects the rank probability skill score (RPSS). The RPSS is a multi-category generalization of the two-category Brier skill score. Richardson (2001) found that finite ensemble size had an adverse effect on the Brier skill score with low-skill regions being more negatively affected by small ensemble size. Using a simple cost-loss decision model, Richardson (2001) noted that changes in ensemble size that cause only modest changes in Brier skill score led to large changes in economic value, particularly for extreme events.

Accurate estimation of tercile probabilities from GCM ensembles does not ensure a skillful

simulation or forecast. Systematic GCM errors may also contribute to error in tercile probabilities, and calibration of model probabilities is needed to account for model deficiencies (Robertson et al. 2004). Generally, we expect that skill will be improved by reducing sampling error in the model-based probabilities that are inputs to the calibration system. More specifically, we investigate the role of model error using a 79-member ensemble of GCM simulations of seasonal precipitation. We examine the impact of reducing sampling error on the skill of the simulations with and without calibration. Additionally, we use the GCM data to assess the importance of some simplifying assumptions used in the calculation of the analytical results, comparing the analytical results with empirical ones obtained by sub-sampling from the ensemble of GCM simulations.

An interesting predictability issue relevant to parametric estimation of tercile probabilities is the question of the roles of the forecast mean and variance in determining predictability. Since predictability is a measure of the difference between forecast and climatological distributions, identifying the parameters that determine predictability identifies the parameters that should be used to estimate tercile probabilities. One approach to this question is to determine the parameters that give the most skillful forecast probabilities (Buizza and Palmer 1998; Atger 1999). Kharin and Zwiers (2003) showed that the Brier skill score of hindcasts of 700 mb temperature and 500 mb height was improved when probabilities were estimated from a Gaussian distribution with constant variance as compared with counting; fitting a Gaussian distribution with time-varying variance gave inferior results. Hamill et al. (2004) used a generalized linear model (GLM; logistic regression) to estimate forecast tercile probabilities of 6-10 day and week-2 surface temperature and precipitation and found that the ensemble variance was not a useful predictor of tercile probabilities. In addition to looking at skill, we examine this question in the perfect model setting by asking whether including ensemble variance in the Gaussian estimate and the GLM estimate reduces sampling error.

The paper is organized as follows. The GCM and observation data are described in section 2. In section 3, we derive some theoretical results about the relative size of the error of the counting

and fitting estimates, and about the effect of sampling error on the ranked probability skill score. The GLM is also introduced and related to Gaussian fitting. In section 4, we compare the analytical results with empirical GCM-based ones and include effects of model error. A summary and conclusions are given in section 5.

## 2. Data

Model simulated precipitation data come from a 79-member ensemble of T42 ECHAM 4.5 GCM (Roeckner et al. 1996) simulations forced with observed SST for the period December 1950 to February 2003. We use seasonal averages of the three month period December through February (DJF), a period when ENSO is a significant source of predictability. We consider all land points between 55S and 70N, including regions whose dry season occurs in DJF and where forecasts are not usually made. While the results here use unprocessed model simulated precipitation, many of the calculations were repeated using Box-Cox transformed data. The Box-Cox transformation

$$x_{BC} = \begin{cases} \lambda^{-1}\left(x^{\lambda} - 1\right) , & \lambda \neq 0 \\ \log x , & \lambda = 0 \end{cases} \tag{1}$$

makes the data approximately Gaussian and depends on the parameter $\lambda$. Positive skewness is the usual non-Gaussian aspect of precipitation. The value of $\lambda$ is found by maximizing the log likelihood function. Figure 1 shows the geographical distribution of the values of $\lambda$ which is an indication of the deviation of the data from Gaussianity; we only allow a few values of $\lambda$, namely 0, 1/4, 1/3, 1/2 and 1. The log function and small values of the exponent tend to be selected in dry regions. This is consistent with Sardeshmukh et al. (2000) who found that monthly precipitation in Reanalysis and in a GCM was significantly non-Gaussian mainly in regions of mean tropospheric descent.

The precipitation observations used to evaluate model skill and to calibrate model output come from the extended New et al. (2000) gridded dataset of monthly precipitation for the period 1950

to 1998, interpolated to the T42 model grid.

## 3. Theoretical considerations

### *a. Variance of the counting estimate*

The *counting estimate* $p_N$ of a tercile probability is the fraction $n/N$ where $N$ is the ensemble size and $n$ is the number of ensemble members in the tercile category. The binomial distribution $P_p(n|N)$, where $p$ is the tercile probability, gives the probability of there being exactly $n$ members in the category. The expected number of members in the tercile category is

$$\langle n \rangle = \sum_{n=0}^{N} n P_p(n|N) = Np\,, \tag{2}$$

where the notation $\langle \cdot \rangle$ denotes expectation. Consequently, the expected value of the counting estimate $p_N$ is the probability $p$, and the counting estimate is unbiased. However, having a limited ensemble size generally causes any single realization of $p_N$ to differ from $p$. The variance of the counting estimate $p_N$ is

$$\langle (p_N - p)^2 \rangle = \sum_{n=0}^{N} \left( \frac{n}{N} - p \right)^2 P_p(n|N) = \frac{1}{N^2} \sum_{n=0}^{N} (n - pN)^2 P_p(n|N) = \frac{1}{N}(1-p)p\,, \tag{3}$$

where we have used the fact that the variance of the binomial distribution is $N(1-p)p$. The relation in (3) implies that the standard deviation of the counting estimate decreases as $N^{-1/2}$, a convergence rate commonly observed in Monte Carlo methods.

Since the counting estimate $p_N$ is not normally distributed or even symmetric for $p \neq 0.5$ (for instance, the distribution of sampling error necessarily has a positive skew when the true probability $p$ is close to zero), it is not immediately apparent whether its variance is a useful measure. However, the binomial distribution becomes approximately normal for large $N$. Figure 2 shows that the standard deviation gives a good estimate of the 16th and 84th percentiles of $p_N$ for $p = 1/3$ and modest values of $N$. In this case, the counting estimate variance is $2/9N$. The

percentiles are obtained by inverting the cumulative distribution function of the sample error. Since the binomial cumulative distribution is discrete, we show the smallest value at which it exceeds 0.16 and 0.84. Figure 2 also shows that for modest sized ensembles ($N > 20$) the standard deviation is fairly insensitive to incremental changes in ensemble size; increasing the ensemble size by a factor of 4 is necessary to reduce the standard deviation by a factor of 2.

The average variance of the counting estimate for a number of forecasts is found by averaging (3) over the values of the probability $p$. The extent to which the forecast probability differs from the climatological value of $1/3$ is an indication of predictability, with larger deviations indicating more predictability. Intuitively, we expect regions and seasons with more predictability to suffer less from sampling error on average, since enhanced predictability implies more reproducibility among ensemble members. In fact, when the forecast distribution is Gaussian with mean $\mu_f$ and variance $\sigma_f$, the variance of the counting estimate of the below-normal category probability is (see the Appendix for details)

$$p - p^2 = \frac{1}{4}\left\{1 - \left[\mathrm{erf}\left(\frac{x_b - \mu_f}{\sqrt{2}\sigma_f}\right)\right]^2\right\}, \tag{4}$$

where $x_b$ is the left tercile boundary. The variance is small when the ensemble mean is large or the ensemble variance is small. Assuming that the forecast variance is constant and averaging (4) over forecasts gives that the average variance is approximately (see the Appendix for details)

$$\left\langle p - p^2 \right\rangle \approx -\frac{0.0421868}{N} + \frac{0.264409}{N\sqrt{1+S^2}} \approx \frac{2}{9N\sqrt{1+S^2}}, \tag{5}$$

where $S^2$ is the usual signal-to-noise ratio (see (A.4); Kleeman and Moore 1999; Sardeshmukh et al. 2000). The signal-to-noise ratio is related to correlation skill with $S/\sqrt{1+S^2}$ being the expected correlation of the ensemble mean with an ensemble member. The relation in (5) has the practical value of providing a simple estimate of the ensemble size needed to to achieve a given level of accuracy for the counting estimate of the tercile probability. This value, like the signal-to-noise ratio, depends on the model, season and region.

*b. Variance of the Gaussian fit estimate*

Fitting a distribution with a few adjustable parameters to the ensemble precipitation is an alternative method of estimating a quantile probability. Here we use a Gaussian distribution with two parameters, mean and variance, for simplicity and because it can be easily generalized to more dimensions (Wilks 2002). The *Gaussian fit estimate* $g_N$ of the tercile probabilities is found by fitting the $N$-member ensemble with a Gaussian distribution and integrating the distribution between the climatological tercile boundaries (Kharin and Zwiers 2003). The Gaussian fit estimate has two sources of error: (i) the non-Gaussianity of the forecast distribution from which the ensemble is sampled and (ii) sampling error in the estimates of mean and variance due to limited ensemble size. The first source of error is problem dependent, and we will quantify its impact empirically for the case of GCM simulated seasonal precipitation. The variance of the Gaussian fit estimate can be quantified analytically for Gaussian distributed variables. When the forecast distribution is Gaussian with mean $\mu_f$ and known variance $\sigma_f$, the variance of the Gaussian fit estimate of the below-normal category probability is approximately (see the Appendix for details)

$$\frac{1}{2\pi N}\exp\left[-\left(\frac{x_b-\mu_f}{\sigma_f}\right)^2\right], \tag{6}$$

where $x_b$ is the left tercile boundary. The average (over forecasts) variance of the Gaussian fit tercile probability is approximately (see Appendix for details)

$$\left\langle (p-g_N)^2\right\rangle \approx \frac{\exp\left(-\frac{1+S^2}{1+2S^2}x_0^2\right)}{2\pi N\sqrt{1+2S^2}}\,. \tag{7}$$

Comparing this value with the counting estimate variance in (5) shows that the Gaussian fit estimate has smaller variance for all values of $S^2$, with its advantage over the counting estimate increasing slightly as the signal-to-noise ratio increases to levels exceeding unity.

When there is no predictability ($S=0$)

$$\left\langle (p-g_N)^2\right\rangle \approx \frac{e^{-x_0^2}}{2\pi N} \approx \frac{0.1322}{N}\,. \tag{8}$$

Comparing (3) and (8), we see that the variance of the Gaussian estimated tercile probability is about 40% smaller than that of the counting estimate if the ensemble distribution is indeed Gaussian with known variance and no signal ($S = 0$). The inverse dependence of the variances on ensemble size means that modest decreases in variance are equivalent to substantial increases in ensemble size. For instance, the variance of a Gaussian fit estimate with ensemble size 24, the simulation ensemble size used for IRI forecast calibration (Robertson et al. 2004), is equivalent to that of a counting estimate with ensemble size 40. The results in (3) and (8) also allow us to compare the variances of counting and Gaussian fit estimates of other quantile probabilities for the case $S = 0$ by appropriately modifying the definition of the category boundary $x_0$. For instance, to estimate the median, $x_0 = 0$, and the variance of the Gaussian estimate is about 36% smaller than that of the counting estimate; in the case of the 10th and 90th percentiles, $x_0 = -1.2816$, and the variance of the Gaussian estimated probability is about 66% smaller than that of the counting estimate. The accuracy of the approximation in (8) for higher quantiles depends on the ensemble size being sufficiently large.

*c. Estimates from generalized linear models*

Generalized linear models (GLMs) offer a parametric estimate of quantile probabilities without the explicit assumption that the ensemble have a Gaussian distribution. GLMs arise in the statistical analysis of the relationship between a response probability $p$, here the tercile probability, and some set of explanatory variables $y_i$, as for instance the ensemble mean and variance (McCullagh and Nelder 1989). Suppose the probability $p$ depends on the response $R$, which is the linear combination

$$R = \sum_i a_i y_i + b\,, \tag{9}$$

of the explanatory variables for some coefficients $a_i$ and a constant term $b$. The response $R$ generally takes on all numerical values while the probability $p$ is bounded between zero and one. The GLM approach introduces a function $g(p)$ that maps the unit interval on the entire real line and studies the model

$$g(p) = R = \sum_i a_i y_i + b\,. \tag{10}$$

The parameters $a_i$ and the constant $b$ are found by maximum likelihood estimation. Here, the GLMs are developed with the ensemble mean (standardized) and ensemble standard deviation as explanatory variables and $p$ given by the counting estimate applied to the same ensemble.

There are a number of commonly used choices for the function $g(p)$ including the logit function which leads to logistic regression (McCullagh and Nelder 1989; Hamill et al. 2004). Here we use the probit function which is the inverse of the normal cumulative distribution function $\Phi$, that is, we define

$$g(p) \equiv \Phi^{-1}(p)\,. \tag{11}$$

Results using the logit function (not shown) are similar since the logistic and probit function are very similar over the interval $0.1 \leq p \leq 0.9$ (McCullagh and Nelder 1989). The assumption of the GLM method is that $g(p)$ is linearly related to the explanatory variables: here, the ensemble mean and standard deviation. When the forecast distribution is Gaussian with constant variance, $g(p)$ is indeed linearly related to the ensemble mean and this assumption is exactly satisfied. To see this, suppose that the forecast ensemble has mean $\mu_f$ and variance $\sigma_f$. Then the probability $p$ of the below-normal category is

$$p = \Phi\left(\frac{x_b - \mu_f}{\sigma_f}\right)\,, \tag{12}$$

where $x_b$ is the left tercile of the climatological distribution, and

$$g(p) = \frac{x_b - \mu_f}{\sigma_f}\,. \tag{13}$$

Therefore, we expect the Gaussian fit and GLM estimates to have similar behavior for Gaussian ensembles with constant variance, with the only differences due to different methods of estimating the mean.

We show an example with synthetic data to give some indication of the robustness of the GLM estimate when the population that the ensemble represents does not have a Gaussian distribution. We take the forecast pdf to be a gamma distribution with shape and scale parameters (2,1). The pdf is asymmetric and has a positive skew (see Fig. 3a). Samples are taken from this distribution and the probability of the below-normal category is estimated by counting, Gaussian fit and GLM; the Gaussian fit assumes constant known variance, and the GLM uses the ensemble mean as an explanatory variable. Interestingly the rms error of both the GLM and Gaussian fit estimates are smaller than that of counting for modest ensemble size (Fig. 3b). As the ensemble size increases further, counting becomes a better estimate than the Gaussian fit. For all ensemble sizes, the performance of the GLM estimate is better than the Gaussian fit.

Other experiments (not shown) compare the counting, Gaussian fit and GLM estimates when the ensemble is Gaussian with non-constant variance. The GLM estimate with ensemble mean and variance as explanatory variables and the 2-parameter Gaussian fit have smaller error than counting and the 1-parameter models (for large enough ensemble size) as predicted.

*d. Ranked probability skill score*

The ranked probability skill score (RPSS; Epstein 1969), a commonly used skill measure for probabilistic forecasts, is also affected by sampling error. The ranked probability score (RPS) is the average integrated squared difference between the forecast and observed cumulative distribution functions and is defined for tercile probabilities to be

$$RPS \equiv \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{3} (F_{i,j} - O_{i,j})^2 \,, \tag{14}$$

where $M$ is the number of forecasts, $F_{i,j}$ ($O_{i,j}$) is the cumulative distribution function of the $i$th forecast (observation) of the $j$th category. The observation "distribution" is defined to be one for the observed category and zero otherwise. This definition means that $F_{i,1} = P_{i,B}$, $F_{i,2} = P_{i,B}+P_{i,N}$ where $P_{i,B}$ ($P_{i,N}$) is the probability of the below-normal (near-normal) category for the $i$th forecast. The terms with the above-normal probabilities ($j = 3$) vanish.

Suppose we consider the expected (with respect to realizations of the observations) RPS for a particular forecast and for simplicity drop the forecast number subscript. Let $\mathcal{O}_B$, $\mathcal{O}_N$, and $\mathcal{O}_A$ be the probabilities that the observation falls into the below-, near- and above-normal categories, respectively. That is,

$$\mathcal{O}_B = \langle O_1 \rangle \ , \quad \mathcal{O}_N = \langle O_2 \rangle \ , \quad \mathcal{O}_A = \langle O_3 \rangle \ , \tag{15}$$

where the expectation is with respect to realizations of the observations. Note that $\mathcal{O}_B$, $\mathcal{O}_N$, and $\mathcal{O}_A$ collectively represent the uncertainty of the climate state, not due to instrument error but due to the limited predictability of the climate system. These quantities are not directly measurable since only a single realization of nature is available. The expected (with respect to the observations) RPS of a particular forecast is the sum of the RPS for each possible category of observation multiplied by its likelihood

$$\begin{aligned} \langle RPS \rangle &= \mathcal{O}_B \left[ (P_B - 1)^2 + (P_B + P_N - 1)^2 \right] + \mathcal{O}_N \left[ P_B^2 + (P_B + P_N - 1)^2 \right] \\ &\quad + \mathcal{O}_A \left[ P_B^2 + (P_B + P_N)^2 \right] \ , \\ &= \mathcal{O}_B \left[ (P_B - 1)^2 + P_A^2 \right] + \mathcal{O}_N \left[ P_B^2 + P_A^2 \right] + \mathcal{O}_A \left[ P_B^2 + (1 - P_A)^2 \right] \ . \end{aligned} \tag{16}$$

Under the perfect model assumption, observations and forecasts are drawn from the same distribution, and the observation and forecast probabilities are equal. Using (16), the perfect model expected RPS (denoted $RPS_{\text{perfect}}$) is

$$\begin{aligned} RPS_{\text{perfect}} &\equiv \mathcal{O}_B \left[ (\mathcal{O}_B - 1)^2 + \mathcal{O}_A^2 \right] + \mathcal{O}_N \left[ \mathcal{O}_B^2 + \mathcal{O}_A^2 \right] + \mathcal{O}_A \left[ \mathcal{O}_B^2 + (1 - \mathcal{O}_A)^2 \right] \\ &= \mathcal{O}_B(1 - \mathcal{O}_B) + \mathcal{O}_A(1 - \mathcal{O}_A) \, . \end{aligned} \tag{17}$$

Note that the expected RPS of a perfect model differs from zero unless the probability of a category is one, or zero, i.e., unless the forecast is deterministic. The quantity $RPS_{\text{perfect}}$ is a perfect model measure of potential probabilistic skill analogous to the signal-to-noise ratio which determines the correlation skill of a model to predict itself. In the decomposition of Murphy (1973), $RPS_{\text{perfect}}$ measures the *uncertainty*. When the forecast distribution is Gaussian, $RPS_{\text{perfect}}$ is simply related to the forecast mean $\mu_f$ and variance $\sigma_f^2$ by

$$RPS_{\text{perfect}} = 2\Phi\left(\frac{x_b - \mu_f}{\sigma_f}\right)\left(1 - \Phi\left(\frac{x_b - \mu_f}{\sigma_f}\right)\right) . \tag{18}$$

The above formula elucidates the empirical relation between probability skill and mean forecast found by Kumar et al. (2001).

Figure 4a shows the time-averaged value of $RPS_{\text{perfect}}$ for the 79-member ECHAM 4.5 GCM simulated precipitation data. This is a perfect model measure of potential probabilistic skill with small values of $RPS_{\text{perfect}}$ showing the GCM has skill in the sense of reproducibility with respect to itself. Skills are highest at low latitudes, consistent with our knowledge that tropical precipitation is most influenced by SST. Perfect model RPS values are close to the no-skill limit of 4/9 in much of the extratropics. The ranked probability skill score (RPSS) is defined using the RPS and a reference forecast defined to have zero skill, here climatology:

$$RPSS = 1 - \frac{RPS}{RPS_{\text{clim}}} , \tag{19}$$

where $RPS_{\text{clim}}$ is the RPS of the climatological forecast. The expected RPS of a climatological forecast is found by substituting $P_B = P_N = P_A = 1/3$ into (16) which gives

$$RPS_{\text{clim}} = \frac{2}{9} + \frac{1}{3}\left(\mathcal{O}_B + \mathcal{O}_A\right) . \tag{20}$$

Figure 4b shows the time-averaged value of $RPSS_{\text{perfect}} \equiv 1 - RPS_{\text{perfect}}/RPS_{\text{clim}}$ for the GCM simulated precipitation data. Even with the perfect model assumption, the RPSS exceeds 0.1 in few regions.

Even in the perfect model setting, the ensemble-estimated and observation probabilities are different due to finite ensemble size. Suppose that $P_B = \mathcal{O}_B + \epsilon_B$ and $P_A = \mathcal{O}_A + \epsilon_A$ where $\epsilon_B$ and $\epsilon_A$ represent error due to finite ensemble size. If each of the forecast probabilities are unbiased and $\langle \epsilon_B \rangle = \langle \epsilon_A \rangle = 0$, then substituting into (16) and averaging over realizations of the ensemble gives

$$\begin{aligned} \langle RPS \rangle &= RPS_{\text{perfect}} + \mathcal{O}_B \left\langle \epsilon_B^2 + \epsilon_A^2 \right\rangle + \mathcal{O}_N \left\langle \epsilon_B^2 + \epsilon_A^2 \right\rangle + \mathcal{O}_A \left\langle \epsilon_B^2 + \epsilon_A^2 \right\rangle \\ &= RPS_{\text{perfect}} + \left\langle \epsilon_B^2 + \epsilon_A^2 \right\rangle . \end{aligned} \tag{21}$$

This means that in the perfect model setting the expected RPS is increased by an amount that depends on the variance of the probability estimate. In particular, if the sampling error is associated with the counting estimate whose variance is given by (3), then

$$\left\langle \epsilon_B^2 + \epsilon_A^2 \right\rangle = \frac{1}{N} \left( \mathcal{O}_B(1 - \mathcal{O}_B) + \mathcal{O}_A(1 - \mathcal{O}_A) \right) , \tag{22}$$

and

$$\langle RPS \rangle = \left( 1 + \frac{1}{N} \right) RPS_{\text{perfect}} . \tag{23}$$

It follows that

$$\begin{aligned} \langle RPSS \rangle &= 1 - \left( 1 + \frac{1}{N} \right) \frac{RPS_{\text{perfect}}}{RPS_{\text{clim}}} , \\ &= \frac{(N+1) RPSS_{\text{perfect}} - 1}{N} . \end{aligned} \tag{24}$$

The relation between RPSS and ensemble size is the same as that for the Brier skill score (Richardson 2001). The relation in (24) quantifies the degradation of RPSS due to sampling error, and combined with (18), provides an analytical expression for the empirical relation between ensemble size, RPSS and mean forecast found in Kumar et al. (2001).

If the tercile probability estimate has variance that differs from that of the counting estimate by some factor $\alpha$, as does, for example, the Gaussian fit estimate, then

$$RPSS = \frac{(N + \alpha) RPSS_{\text{perfect}} - \alpha}{N} , \tag{25}$$

where degradation of the RPSS is reduced for $\alpha < 1$.

## 4. Estimates of GCM simulated seasonal precipitation tercile probability

### *a. Variance of the counting estimate*

The average variance of the counting estimate in (5) was derived assuming Gaussian distributions. To see how well this approximation describes the behavior of GCM simulated seasonal precipitation totals, we compare the average counting estimate variance in (5) to that computed by sub-sampling from a large ensemble of GCM simulations. We use the fact that the average squared difference of two independent counting estimates is twice the variance. More specifically, we select two independent samples of size $N$ (without replacement) from the ensemble of GCM simulations and compute two counting estimate probabilities denoted $p_N$ and $p'_N$; the ensemble size of 79 and independence requirement limits the maximum value of $N$ to 39. The expected value of the square of the difference between the two counting estimates $p_N$ and $p'_N$ is twice the variance of the counting estimate since

$$\left\langle (p_N - p'_N)^2 \right\rangle = \left\langle ((p_N - p) + (p - p'_N))^2 \right\rangle = 2 \left\langle (p_N - p)^2 \right\rangle , \tag{26}$$

where we use the fact that the sampling errors $(p_N - p)$ and $(p - p'_N)$ are uncorrelated. The averages in (26) are with respect to time and realizations (1000) of the two independent samples.

We expect especially close agreement between the sub-sampling calculations and the analytical results of (5) in regions where there is little predictability and the signal-to-noise ratio $S^2$ is small, since, for $S^2 = 0$, the analytical result is exact. In regions where the signal-to-noise ratio is not zero, though generally fairly small, we expect that the average counting variance still decreases as $1/N$ but there is no guarantee that the Gaussian approximation will provide a useful description of the actual behavior of the GCM data. However, Fig. 5 shows that in the land gridpoint average, the variance of the counting estimate is very well described by the analytical result in (5), with the difference from the analytical result being on the order of a few percent for the below-normal category probability and less than one percent for the above-normal category probability.

Figure 6a shows the spatial variation of the convergence factor $-0.0421868+0.264409/\sqrt{1+S^2}$ appearing in (5). This factor can be interpreted as the variance of the counting estimate based on a single member ensemble; the counting estimate standard deviation for ensemble of size $N$ is obtained by dividing by $\sqrt{N}$. This convergence factor can also be obtained empirically from sub-samples of varying size. The difference between the theoretical factor and the empirical estimate is mostly on the order of a few percent (see Figs. 6b and 6c).

We now use sub-sampling of the GCM simulated precipitation data to compare the three estimation methods–counting, Gaussian fit and GLM–discussed in the previous section. Since the Gaussian fit and GLM estimators may be biased, it is not sufficient to compute their variance. The *error* variance of the estimators must be computed. The error is not known exactly because the true probability is not known exactly. Therefore each method is compared to a common baseline as follows. Each method is applied to an ensemble of size $N$ ($N = 5, 10, 20, 30, 39$) to produce an estimate $q_N$. This estimate is then compared to the counting estimate $p_{40}$ computed from an independent set of 40 ensemble members. This counting estimate $p_{40}$ serves as a common unbiased baseline. The variance of the difference of these two estimates has contributions from the $N$-member estimate $q_N$ and the 40-member counting estimate. The variance of the difference can be decomposed into error variance contributions from $q_N$ and $p_{40}$:

$$\begin{aligned}\langle (q_N - p_{40})^2 \rangle &= \langle (q_N - p + p - p_{40})^2 \rangle \\ &= \langle (q_N - p)^2 \rangle + \langle (p - p_{40})^2 \rangle \\ &\approx \langle (q_N - p)^2 \rangle - \frac{0.0421868}{40} + \frac{0.264409}{40\sqrt{1+S^2}}\,,\end{aligned} \tag{27}$$

where the theoretical estimate of the variance of $p_{40}$ is used. Therefore, the error variance of the estimate $q_N$ is:

$$\langle (q_N - p)^2 \rangle \approx \langle (q_N - p_{40})^2 \rangle + \frac{0.0421868}{40} - \frac{0.264409}{40\sqrt{1+S^2}}\,. \tag{28}$$

All results for the estimate error variance are presented in terms of $\langle (q_N - p)^2 \rangle$ rather than $\langle (q_N -$

$p_{40})^2\rangle$ so as to give a sense of the magnitude of the sampling error rather than the difference with the baseline estimate. Results are averaged over time and realizations (100) of the $N$-member estimate and the 40-member counting estimate.

We begin by examining the land gridpoint average of the sampling error of the three methods. Figure 7a shows the gridpoint averaged rms error of the tercile probability estimates as a function of ensemble size. The variance of the counting estimate is well-described by theory (Fig. 7a) and is larger than that of the parametric estimates. The one-parameter GLM and constant variance Gaussian fit have similar rms error for larger ensemble sizes; the GLM estimate is slightly better for very small ensemble sizes. While the magnitude of the error reduction due to using the parametric estimates is modest, the savings in computational cost compared to the equivalent ensemble size is significant. The single parameter estimates, that is, the constant variance Gaussian fit and the GLM based on the ensemble mean, have smaller rms error than the estimates based on ensemble mean and variance (Fig. 7b). The advantage of the single parameter estimates is greatest for smaller ensemble sizes. This result is important because it shows that attempting to account for changes in variance, even in the perfect model setting, does not improve estimates of the tercile probabilities for the range of ensemble sizes considered here (Kharin and Zwiers 2003). The sensitivity of the tercile probabilities to changes in variance is, of course, problem specific.

Figure 8 shows the spatial features of the rms error of the below-normal tercile probability estimates for ensemble size 20. Using a Gaussian with constant variance or a GLM based on the ensemble mean has error that is, on average, less than counting; the average performances of the Gaussian fit and the GLM are similar. In a few dry regions, especially in Africa, the error from the parametric estimates is larger. This problem with the parametric estimates in the dry regions is reduced when a Box-Cox transformation is applied to the data (not shown), and overall error levels are slightly reduced as well. The spatial features of rms error when the variance of the Gaussian is estimated and when the mean and standard deviation are used in the GLM are similar to those in

Fig. 8, but the overall error levels are slightly higher.

*b. RPSS*

Having evaluated the three probability estimation methods in the perfect model setting where we asked how closely they match the probabilities from a large ensemble, we now use the RPSS to compare the GCM estimated probabilities with observations. We expect the reduction in sampling error to result in improved RPSS but we cannot know beforehand the extent to which model error confounds or offsets the reduction in sampling error. Figure 9 shows maps of RPSS for ensemble size 20 for the counting, Gaussian fit and GLM estimates. The results are averaged over 100 random selections of the 20-member ensemble from the full 79-member ensemble. The overall skill of the Gaussian fit and GLM estimate is similar and both are generally larger than that of the counting estimate.

Figure 10 shows the fraction of points with positive RPSS as a function of ensemble size. Again results are averaged over 100 random draws of each ensemble size except for $N = 79$ when the entire ensemble is used. The parametrically estimated probabilities lead to more grid points with positive RPSS . The Gaussian fit and GLM have similar skill levels with the GLM estimate having larger RPSS for the smallest ensemble sizes, and the Gaussian fit being slightly better for larger ensemble sizes. It is useful to interpret the increases in RPSS statistics in terms of effective ensemble size. For instance, applying the Gaussian fit estimator to a 24-member ensemble give RPSS statistics that are on average comparable to those of the counting estimator applied to a ensemble of size about 39. Although all methods show improvement as ensemble size increases, it is interesting to ask to what extent the improvement in RPSS due to increasing ensemble size predicted by (24) is impacted by the presence of model error. For a realistic approximation of the RPSS in the limit of infinite ensemble size, we compute the RPSS for $N = 1$ and solve (24) for

$RPSS_{\text{perfect}}$; we expect that in this case sampling error dominates model error and the relation in (24) holds approximately. Then we use (24) to compute the gridpoint averaged RPSS for other values of $N$; the theory curve in Fig. 10 shows these values. In the absence of model error, the count and theory curves of RPSS in Fig. 10 would be the same. However, we see that the effect of model error is such that the curves are close for $N = 5$ and $N = 10$, and diverge for larger ensemble sizes with the actual increase in RPSS being lower than that predicted by (24).

The presence of model error means that some calibration of the model output with observations is needed. The GCM ensemble tends to be over-confident, and calibration tempers this. To see if reducing sampling error still has an noticeable impact after calibration, we use a simple version of Bayesian weighting (Rajagopalan et al. 2002; Robertson et al. 2004). In the method, the calibrated probability is a weighted average of the GCM probability and the climatology probability $(1/3)$. The weights are chosen to maximize the likelihood of the observations. There is cross-validation in the sense that the weights are computed with a particular ensemble of size $N$ and the RPSS is computed by applying those weights to a different ensemble of the same size and then comparing the result with observations. The calibrated counting-estimated probabilities still have slightly negative RPSS in some areas (Fig. 11a) but the overall amount of positive RPSS is increased compared to the uncalibrated simulations (compare with Fig. 9a); the ensemble size is 20 and results are averaged over 100 realizations. The calibrated Gaussian and GLM probabilities have modestly higher overall RPSS than the calibrated counting estimates with noticeable improvement in skillful areas like Southern Africa (Figs. 11b,c). We note that a simpler calibration method based on a Gaussian fit with the variance determined by the correlation between ensemble mean and observations, as in Tippett et al. (2005), rather than ensemble spread, performs nearly as well as the Gaussian fit with Bayesian calibration.

It is interesting to look at examples of the probabilities given by the counting and Gaussian fit estimate to see how the spatial distributions of probabilities may differ in appearance. Figure

12 shows uncalibrated tercile probabilities from DJF 1996 (ENSO-neutral) and 1998 (strong El Niño). Counting and Gaussian probabilities appear similar, with Gaussian probabilities appearing spatially smoother.

## 5. Summary and conclusions

Here we have explored how the accuracy of tercile probability estimates are related to ensemble size and the chosen probability estimation technique. The counting estimate, which uses the fraction of ensemble members that fall in the tercile category, is attractive since it places no restrictions on the ensemble distribution and is simple. The error variance of the counting estimate is a function of the ensemble size and tercile category probability. For Gaussian variables, the variance of the counting estimate depends on the ensemble size, ensemble mean and variance; the average variance depends on ensemble size and the signal-to-noise ratio. The Gaussian fit estimate computes tercile probabilities from a Gaussian distribution with parameters estimated from the forecast ensemble. The variance of the Gaussian fit tercile probabilities is also a function of the ensemble size and the ensemble mean and variance, and the average variance depends on ensemble size and the signal-to-noise ratio. When the variables are indeed Gaussian, the variance of the Gaussian fit estimate is the error variance and is smaller than that of the counting estimate, by approximately 40% in the limit of small signal. The advantage of the Gaussian fit over the counting estimate is equivalent to fairly substantial increases in ensemble size. However, this advantage depends on the forecast distribution being well described by a Gaussian distribution. Generalized linear models (GLMs) provide a parametric method of estimating the tercile probabilities using a nonlinear regression with the ensemble mean and possibly the ensemble variance as predictors. The GLM estimator does not explicitly assume a distribution but, as implemented here, is equivalent to the Gaussian fit in some circumstances.

The accuracy of the tercile probability estimates affects probability forecast skill measures such as the commonly used ranked probability skill score (RPSS). Reducing the variance of the tercile probability estimate increases the RPSS. We examined this connection in the perfect model setting used extensively in predictability studies in which the "observations" are assumed to be indistinguishable from an arbitrary ensemble member. We find the expected RPSS in terms of the above- and below-normal tercile probabilities and, for Gaussian variables, in terms of the ensemble mean and variance. Finite ensemble size degrades the expected RPSS, conceptually similar to the way that finite ensemble size reduces the expected correlation (Sardeshmukh et al. 2000; Richardson 2001).

Many of the analytical results are obtained assuming that the ensemble variables have a Gaussian distribution. We test the robustness of these findings using simulated seasonal precipitation from an ensemble of GCM integrations forced by observed SST, sub-sampling from the full ensemble to estimate sampling error. We find that the theoretical results give a good description of the average variance of the counting estimate, particularly in a spatially averaged sense. This means that the theoretical scalings can be used in practice to understand how sampling error depends on ensemble size and level of predictability. Although the GCM simulated precipitation departs somewhat from being Gaussian, the Gaussian fit estimate had smaller error than the counting estimate. The behavior of the GLM estimate is similar to that of the Gaussian fit estimate. The parametric estimators based on ensemble mean had the best performance; adding ensemble variance as a parameter did not reduce error. This means that with the moderate ensemble sizes typically used, differences between the forecast tercile probabilities and the equal-odds probabilities are due mainly to shifts of the forecast mean away from its climatological value rather than due to changes in variance. Since differences between the forecast tercile probabilities and the equal-odds probabilities are a measure of predictability, this result means that predictability in the GCM is due to changes in ensemble mean rather than changes in spread. This result is consistent

with Tippett et al. (2004) who found that differences between forecast and climatological GCM seasonal precipitation distributions as measured by relative entropy were primarily due to changes in the mean rather than changes in the variance.

The reduced sampling error of the Gaussian fit and GLM translates into better simulation skill when the tercile probabilities are compared to actual observations. Examining the dependence of the RPSS on ensemble size shows that although RPSS increases with ensemble size, model error limits the rate of improvement compared to the ideal case. Calibration improves RPSS, regardless of the probability estimator used. However, estimators with larger sampling error retain their disadvantage in RPSS even after calibration. The application of the Gaussian fit estimator to specific years shows that the parametric fit achieves its advantages while also producing probabilities that are spatially smoother than those estimated by counting.

In summary, our main conclusion is that carefully applied parametric estimators provide noticeably more accurate tercile probabilities than do counting estimates. This conclusion is completely rigorous for variables with Gaussian statistics. We find that for variables that deviate modestly from Gaussianity, such as seasonal precipitation totals, Gaussian fit methods offer tercile probability accuracy at least equivalent to that of counting estimates but at substantially reduced cost in terms of ensemble size. More substantial deviation from Gaussianity may be treated by transforming the data or using the related GLM approach.

*Acknowledgements*. We thank Benno Blumenthal for the IRI Data Library. Computer resources for this work were provided in part by the NCAR CSL. GCM integrations were performed by David DeWitt, Shuhua Li and Lisa Goddard. Comments from two anonymous reviewers improved greatly improved the clarity of this manuscript. IRI is supported by its sponsors and NOAA Office of Global Programs Grant number NA07GP0213. The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its sub-agencies

## APPENDIX

## Error in estimating tercile probabilities

*a. Derivation of the counting variance*

As shown in (3), the variance of the counting estimate $p_N$ is $(p - p^2)/N$. Suppose the forecast precipitation anomaly has a Gaussian distribution with mean $\mu_f$ and variance $\sigma_f^2$. The probability $p$ of the below-normal category is

$$p = \Phi\left(\frac{x_b - \mu_f}{\sigma_f}\right) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x_b - \mu_f}{\sqrt{2}\sigma_f}\right)\right), \tag{A.1}$$

where $\Phi$ is the normal cumulative distribution function, $\mathrm{erf}$ denotes the error function and $x_b$ is the left tercile boundary of the climatological distribution. In this case, the counting estimate variance depends on the forecast mean and variance through

$$p - p^2 = \frac{1}{4}\left\{1 - \left[\mathrm{erf}\left(\frac{x_b - \mu_f}{\sqrt{2}\sigma_f}\right)\right]^2\right\}. \tag{A.2}$$

Similar relations hold for the above-normal category.

Suppose that the precipitation anomaly $x$ is joint-normally distributed with mean zero and variance $\sigma_x^2$. In this case, the left tercile boundary $x_b$ of the climatological pdf is $\sigma_x x_0$ where $x_0 = \Phi^{-1}(1/3) \approx -0.4307$. Averaging $x^2$ over all forecasts gives that

$$\left\langle x^2 \right\rangle = \sigma_x^2 = \left\langle \mu_f^2 \right\rangle + \sigma_f^2, \tag{A.3}$$

which decomposes the climatological variance $\sigma_x^2$ into signal and noise contributions. We denote the signal variance $\left\langle \mu_f^2 \right\rangle$ by $\sigma_s^2$ and define the signal-to-noise ratio by

$$S^2 \equiv \frac{\sigma_s^2}{\sigma_f^2}. \tag{A.4}$$

Taking the average of (A.2) with respect to forecasts gives

$$\left\langle p - p^2 \right\rangle = \frac{1}{\sqrt{2\pi}\sigma_s}\int_{-\infty}^{\infty}\frac{1}{4}\left\{1 - \left[\mathrm{erf}\left(\frac{x_b - \mu_f}{\sqrt{2}\sigma_f}\right)\right]^2\right\}\exp\left(-\frac{-\mu_f^2}{2\sigma_s^2}\right)\, d\mu_f\,. \tag{A.5}$$

We now make the change of variable $\mu = \sigma_f \mu_f$, and use the fact that $x_b/\sigma_f = x_0\sqrt{1+S^2}$, to obtain

$$\langle p - p^2 \rangle = \frac{1}{\sqrt{2\pi}S} \int_{-\infty}^{\infty} \frac{1}{4} \left\{ 1 - \left[ \mathrm{erf}\left( \frac{x_0\sqrt{1+S^2} - \mu}{\sqrt{2}} \right) \right]^2 \right\} \exp\left( -\frac{-\mu^2}{2S^2} \right) \, d\mu \,. \tag{A.6}$$

From the form of (A.6) we see that the average variance $\langle p - p^2 \rangle$ is a function of the signal-to-noise ratio $S^2$. We express this dependence using a new parameter $g \equiv (1+S^2)^{-1/2}$,

$$\langle p - p^2 \rangle = \frac{g}{4\sqrt{2\pi}\sqrt{g^2-1}} \int_{-\infty}^{\infty} \left\{ 1 - \left[ \mathrm{erf}\left( \frac{x_0/g - \mu}{\sqrt{2}} \right) \right]^2 \right\} \exp\left( -\frac{-g^2\mu^2}{2(g^2-1)} \right) \, d\mu \,. \tag{A.7}$$

To approximate the dependence of the average variance on the signal-to-noise ratio, we perform a series expansion about $g = 1$ corresponding to the signal-to-noise ratio $S^2$ being zero. The first term is found from

$$\langle p - p^2 \rangle_{g=1} = \frac{2}{9} \,, \tag{A.8}$$

and then numerical computation gives that

$$\frac{d}{dg} \langle p - p^2 \rangle_{g=1} = 0.264409 \,. \tag{A.9}$$

An approximation of $\langle p - p^2 \rangle$ is

$$\langle p - p^2 \rangle = \frac{2}{9} + 0.264409(g-1) + \mathcal{O}(g-1)^2 \tag{A.10}$$

or in terms of the signal-to-noise ratio

$$\frac{\langle p - p^2 \rangle}{N} \approx -\frac{0.0421868}{N} + \frac{0.264409}{N\sqrt{1+S^2}} \,. \tag{A.11}$$

This approximation is valid for small values of $S^2$. However, we expect that the variance vanishes as $S^2$ becomes large, and (A.11) does not show this behavior. A remedy is to add a higher order term, in which case

$$\langle p - p^2 \rangle \approx \frac{2}{9} + 0.264409(g-1) + 0.0421868(g-1)^2 = \frac{0.180035}{\sqrt{1+S^2}} + \frac{0.0421868}{1+S^2} \,. \tag{A.12}$$

Since $S^2$ is fairly small for seasonal forecasts, we will use the approximation in (A.11).

*b. Error in estimating tercile probabilities by Gaussian fitting*

Suppose the distributions are indeed Gaussian. We fit the $N$-member forecast ensemble with a Gaussian distribution, using its sample mean $m_f$ and sample variance $s_f^2$ defined by

$$\begin{aligned} m_f &\equiv \frac{1}{N}\sum_{i=1}^{N} x_i \,, \\ s_f^2 &\equiv \frac{1}{N-1}\sum_{i=1}^{N} (x_i - m_f)^2 \,, \end{aligned} \tag{A.13}$$

where $x_i$ denotes the members of the ensemble. Based on this information and using (A.1), the Gaussian fit estimate $g_N$ of the probability of the below-normal category is

$$g_N = \Phi\left(\frac{x_b - m_f}{\sqrt{2}s_f}\right) = \frac{1}{2}\left(1 - \mathrm{erf}\left(\frac{x_b - m_f}{\sqrt{2}s_f}\right)\right) . \tag{A.14}$$

The squared error of the Gaussian fit probability estimate is

$$(g_N - p)^2 = \left\{\frac{1}{2}\left(1 - \mathrm{erf}\left(\frac{x_b - m_f}{\sqrt{2}s_f}\right)\right) - \frac{1}{2}\left(1 - \mathrm{erf}\left(\frac{x_b - \mu_f}{\sqrt{2}\sigma_f}\right)\right)\right\}^2 . \tag{A.15}$$

The error of the Gaussian fit probability estimate is due to error in the sample estimates of the forecast mean and variance.

If there is no predictability and the forecast mean $\mu_f$ is zero, the true tercile probability is 1/3. In this case, the squared error of the Gaussian fit estimate is

$$\left(\frac{1}{3} - \frac{1}{2}\left(1 - \mathrm{erf}\left(\frac{x_b - m_f}{\sqrt{2}s_f}\right)\right)\right)^2 = \frac{m_f^2 e^{-x_0^2}}{2s_f^2\pi} + \mathcal{O}(m_f^3) \,, \tag{A.16}$$

where we have made a Maclaurin expansion in $m_f$. The term $\mathcal{O}(m_f^3)$ can be neglected in what follows for sufficiently large ensemble size $N$; neglecting the higher order terms leads to an underestimate in the final result of about 3.6% for $N = 10$. The quantity $\sqrt{N}m_f/s_f$ has a $t$-distribution with $N-1$ degrees of freedom, and so its variance is $(N-1)/(N-3)$. Therefore the average (over realizations of the ensemble) variance of the Gaussian fit tercile probability is

$$\left\langle \frac{m_f^2 e^{-x_0^2}}{2s_f^2\pi} \right\rangle = \frac{e^{-x_0^2}}{2\pi N}\frac{N-1}{N-3} \approx \frac{0.1322}{N}\frac{N-1}{N-3} \,. \tag{A.17}$$

If the forecast mean is zero for all forecasts, then the forecast variance $\sigma_f^2$ is equal to the climatological variance $\sigma_x^2$ and does not have to be estimated from the ensemble. In that case, the average (over forecasts) variance of the Gaussian fit tercile probability is

$$\left\langle \frac{m_f^2 e^{-x_0^2}}{2\sigma_x^2 \pi} \right\rangle = \frac{e^{-x_0^2}}{2\pi N} \approx \frac{0.1322}{N}\,, \tag{A.18}$$

since $\langle m_f^2 \rangle = \sigma_x^2 / N$.

On the other hand, suppose that the forecast mean is not identically zero, but the forecast variance $\sigma_f$ is constant and known. The squared error of the Gaussian fit probability estimate is

$$(g_N - p)^2 = \left\{ \frac{1}{2}\left(1 - \mathrm{erf}\left(\frac{x_b - m_f}{\sqrt{2}\sigma_f}\right)\right) - \frac{1}{2}\left(1 - \mathrm{erf}\left(\frac{x_b - \mu_f}{\sqrt{2}\sigma_f}\right)\right)\right\}^2 . \tag{A.19}$$

The error is due entirely to the error in estimating the mean. Expanding this expression in a Taylor series in in powers of $(m_f - \mu_f)$ about $m_f = \mu_f$ gives that the squared error is

$$(g_N - p)^2 = \frac{(m_f - \mu_f)^2}{2\pi\sigma_f^2} \exp\left[ -\left(\frac{x_b - \mu_f}{\sigma_f}\right)^2 \right] + \mathcal{O}(\mu_f - m_f)^3\,. \tag{A.20}$$

We now take the expectation of the leading order term in (A.20) with respect to realizations of the ensemble. Since the variance of the estimate of the mean is $\langle (\mu_f - m_f)^2 \rangle = \sigma_f^2 / N$, the squared error as function of forecast mean is

$$\frac{1}{2\pi N} \exp\left[ -\left(\frac{x_b - \mu_f}{\sigma_f}\right)^2 \right] . \tag{A.21}$$

Averaging over forecasts gives that the average variance of the Gaussian fit tercile probability is

$$\begin{aligned} \left\langle (g_N - p)^2 \right\rangle &= \frac{1}{(2\pi)^{3/2}\sigma_f S N} \int_{-\infty}^{\infty} \exp\left[ -\left(\frac{x_b - \mu_f}{\sigma_f}\right)^2 \right] \exp\left( -\frac{\mu_f^2}{\sigma_s^2} \right) d\mu_f \\ &= \frac{\exp\left( -\frac{1+S^2}{1+2S^2} x_0^2 \right)}{2\pi N \sqrt{1+2S^2}}\,, \end{aligned} \tag{A.22}$$

where we use the fact that $x_b^2 / \sigma_f^2 = x_0^2 (1 + S^2)$.

REFERENCES


Atger, F., 1999: The skill of ensemble predition systems. *Mon. Wea. Rev.*, **127**, 1941–1953.

Barnston, A. G., S. J. Mason, L. Goddard, D. G. Dewitt, and S. E. Zebiak, 2003: Multimodel Ensembling in Seasonal Climate Forecasting at IRI. *BAMS*, **84**, 1783–1796.

Buizza, R., and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126**, 2503–2518.

DelSole, T., 2004: Predictability and Information Theory Part I: Measures of Predictability. *J. Atmos. Sci.*, **61**, 2425–2440.

DelSole, T., and M. K. Tippett, 2006: Predictability, information theory, and stochastic models. *Rev. Geophys.* submitted.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8**, 985–987.

Hamill, T. H., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447.

Kharin, V. V., and F. W. Zwiers, 2003: Improved seasonal probability forecasts. *J. Climate*, **16**, 1684–1701.

Kleeman, R., 2002: Measuring dynamical prediction utility using relative entropy. *J. Atmos. Sci.*, **59**, 2057–2072.

Kleeman, R., and A. M. Moore, 1999: A new method for determining the reliability of dynamical ENSO predictions. *Mon. Wea. Rev.*, **127**, 694–705.

Kumar, A., A. G. Barnston, and M. P. Hoerling, 2001: Seasonal predictions, probabilistic verifications, and ensemble size. *J. Climate*, **14**, 1671–1676.

McCullagh, P., and J. A. Nelder, 1989: *Generalized Linear Models*. Chapman and Hall, London.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600.

New, M. G., M. Hulme, and P. D. Jones, 2000: Representing 20th century space-time climate variability. II: Development of 1901-1996 monthly terrestrial climate fields. *J. Climate*, **13**, 2217–2238.

Rajagopalan, B., U. Lall, and S. E. Zebiak, 2002: Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. *Mon. Wea. Rev.*, **130**, 1792–1811.

Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.*, **127**, 2473–2489.

Robertson, A. W., U. Lall, S. E. Zebiak, and L. Goddard, 2004: Improved combination of multiple atmospheric gcm ensembles for seasonal prediction. *Mon. Wea. Rev.*, **132**, 2732–2744. doi: 10.1175/MWR2818.1.

Roeckner, E., K. Arpe, L. Bengtsson, M. Christoph, M. Claussen, L. Dümenil, M. Esch, M. Giorgetta, U. Schlese, and U. Schulzweida, 1996: The atmospheric general circulation model ECHAM-4: Model description and simulation of present-day climate. Technical Report 218, Max-Planck Institute for Meteorology, Hamburg, Germany. 90 pp.

Sardeshmukh, P. D., G. P. Compo, and C. Penland, 2000: Changes of probability associated with El Niño. *J. Climate*, **13**, 4268–4286.

Tippett, M. K., R. Kleeman, and Y. Tang, 2004: Measuring the potential utility of seasonal climate predictions. *Geophys. Res. Lett.*, **31**, L22 201. doi:10.1029/2004GL021575.

Tippett, M. K., L. Goddard, and A. G. Barnston, 2005: Statistical-Dynamical Seasonal Forecasts of Central Southwest Asia winter precipitation. *J. Climate*, **18**, 1831–1843.

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

**List of Figures**

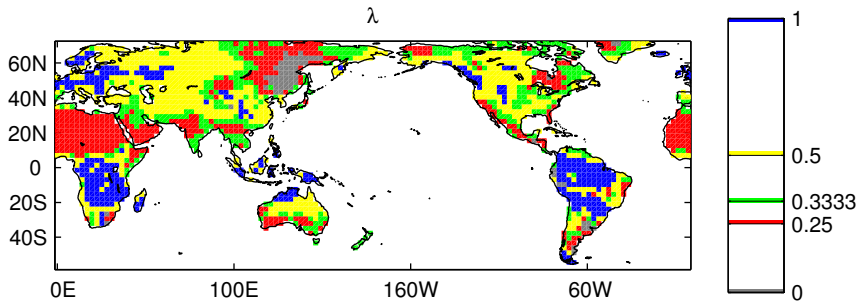



Figure 1. Spatial distribution of $\lambda$ appearing the Box-Cox transformation of Eq. 1.

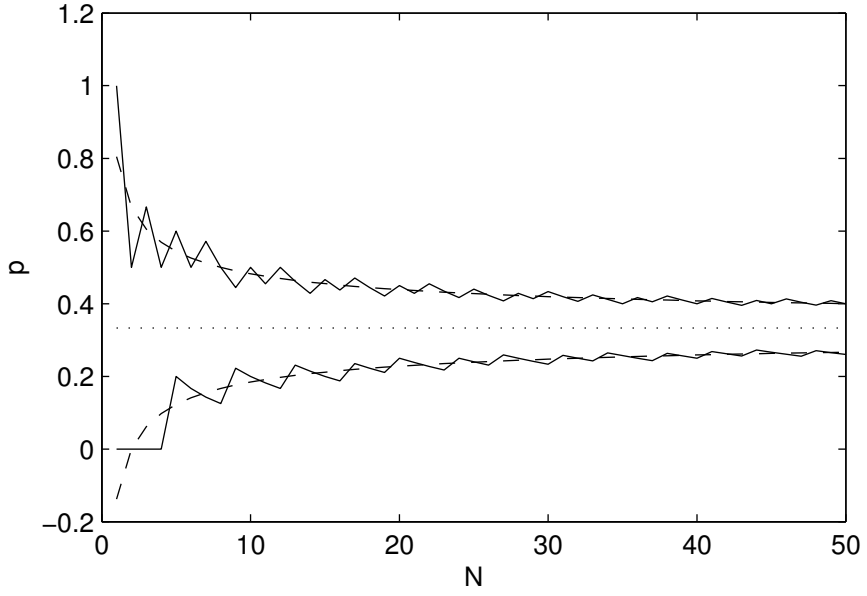



Figure 2. 16th and 84th percentiles (see text for details) of the counting estimate $p_N$ (solid lines) and $p$ plus and minus the standard deviation of the estimate $p_N$ (dashed lines) for $p = 1/3$ (dotted line).

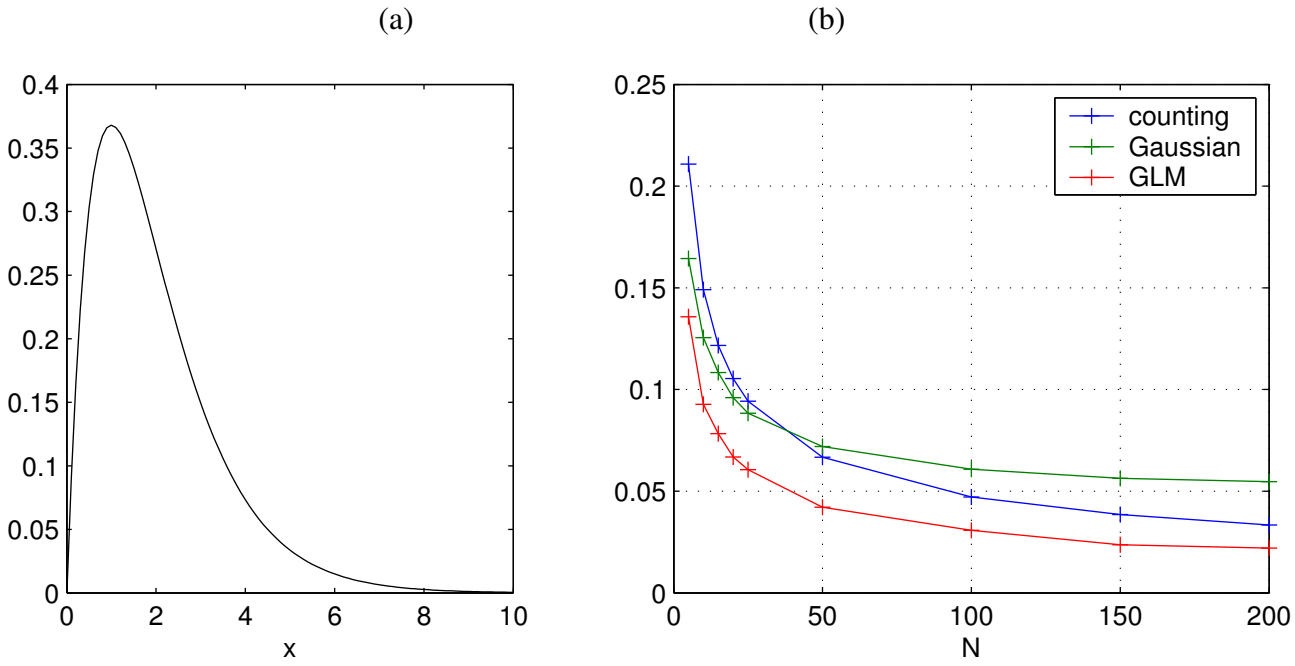



Figure 3. The (a) gamma distribution with shape and scale parameters (2,1), respectively and the (b) rms error as a function of ensemble size $N$ for the counting, Gaussian fit and GLM tercile probability estimates.

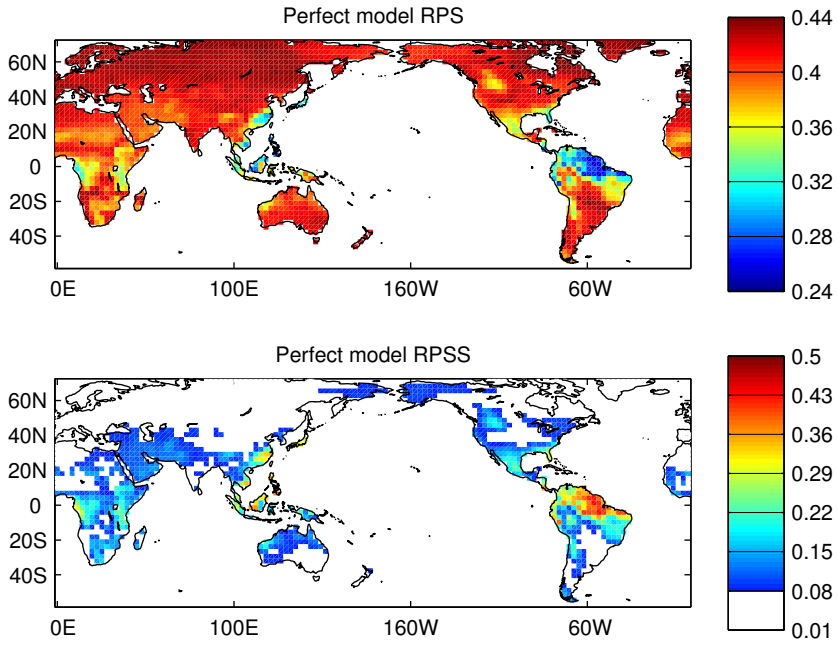



Figure 4. Perfect model measures of potential probability forecast skill (a) $RPS_{\text{perfect}}$ and (b) $RPSS_{\text{perfect}}$ for DJF precipitation.

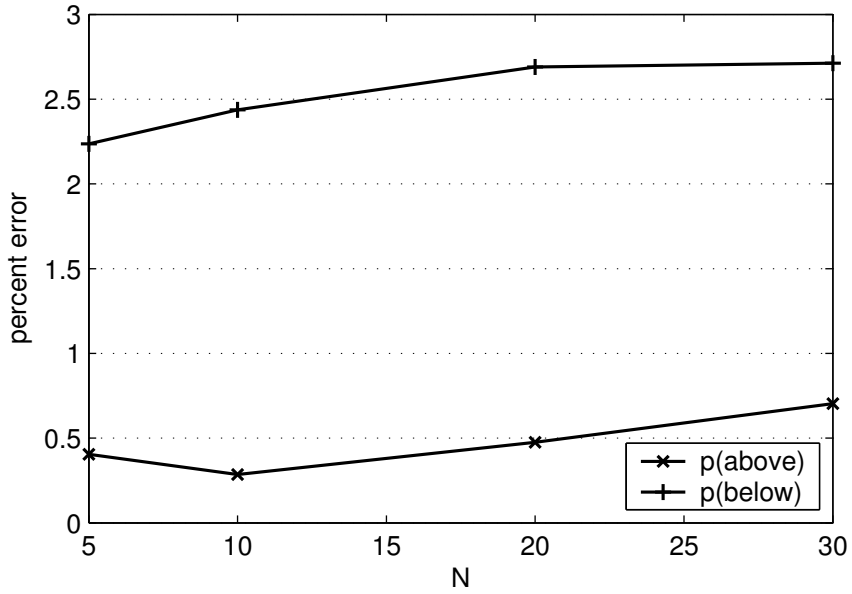



Figure 5. Percent difference between the gridpoint-average of the theoretical and empirically estimated variance of the tercile probability estimate for the below-normal and above-normal categories.

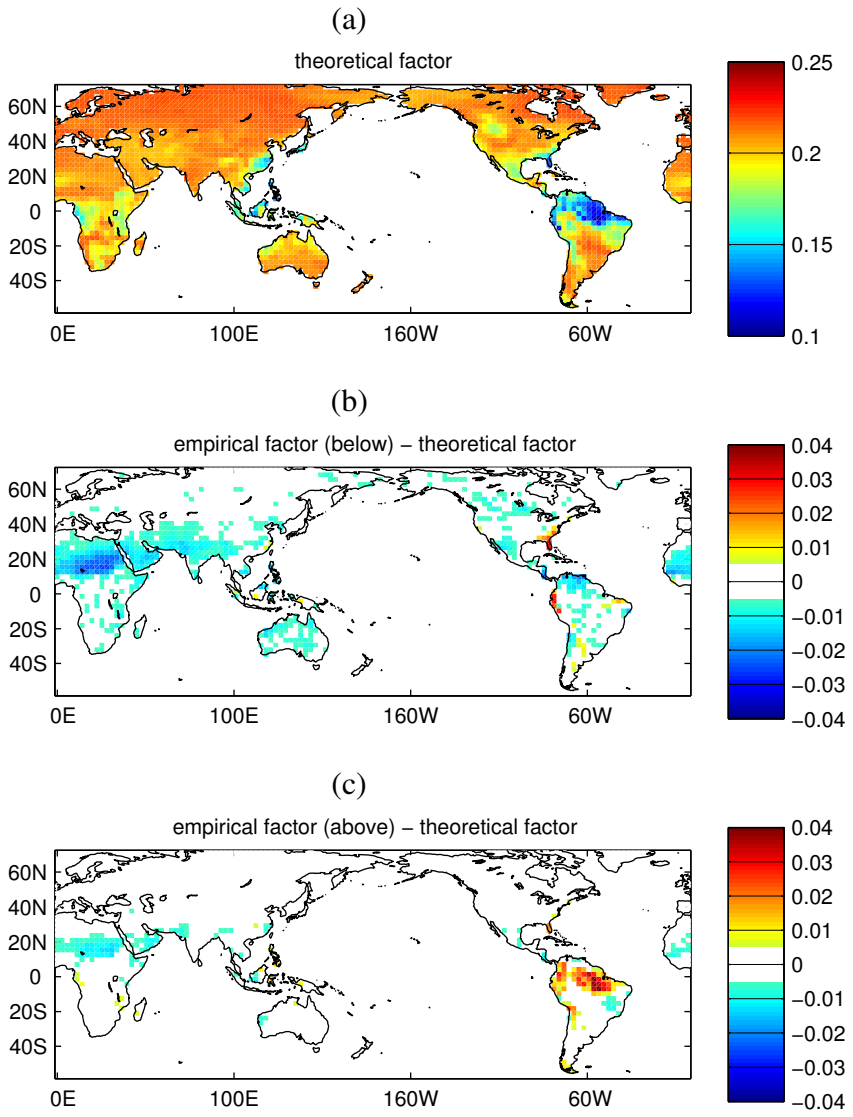
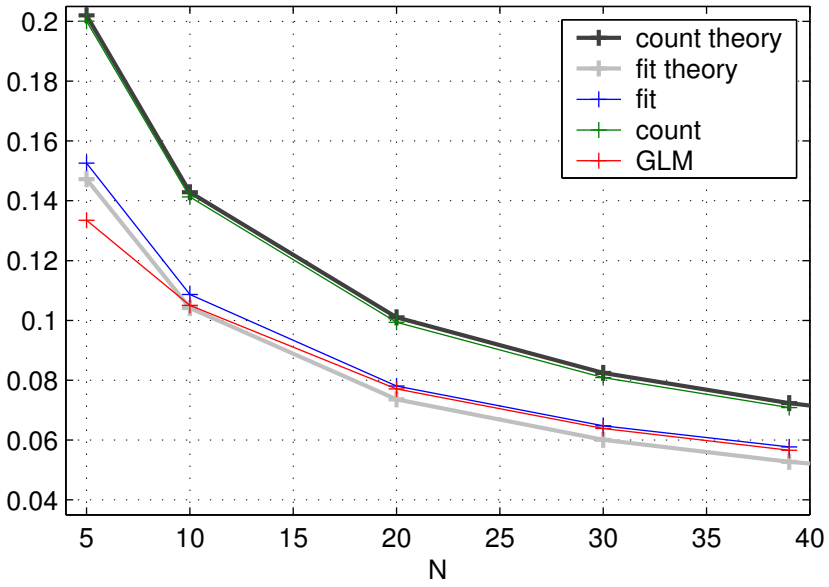



Figure 6. (a) Spatial variation of the convergence factor $-0.0421868 + 0.264409/\sqrt{1+S^2}$. Difference of the theoretical convergence factor with the sub-sampled estimates from the (b) below-normal and (c) above-normal categories.

(a)




(b)

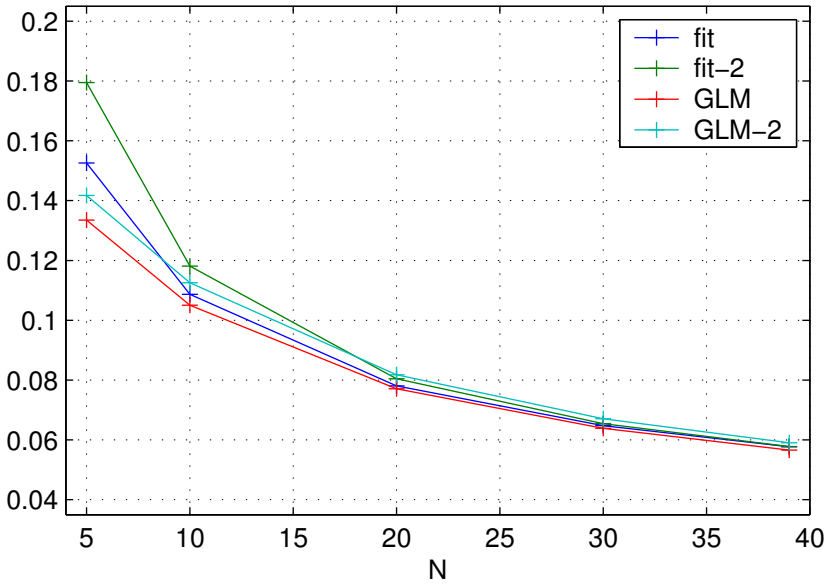



Figure 7. RMS error of the below-normal probability as a function of ensemble size $N$ for the (a) 1-parameter and (b) 2-parameter estimates. The gray curves in panel (a) are the theoretical error levels for the counting and Gaussian fit methods. Fit-2 (GLM-2) denotes the two-parameter Gaussian (GLM) method.

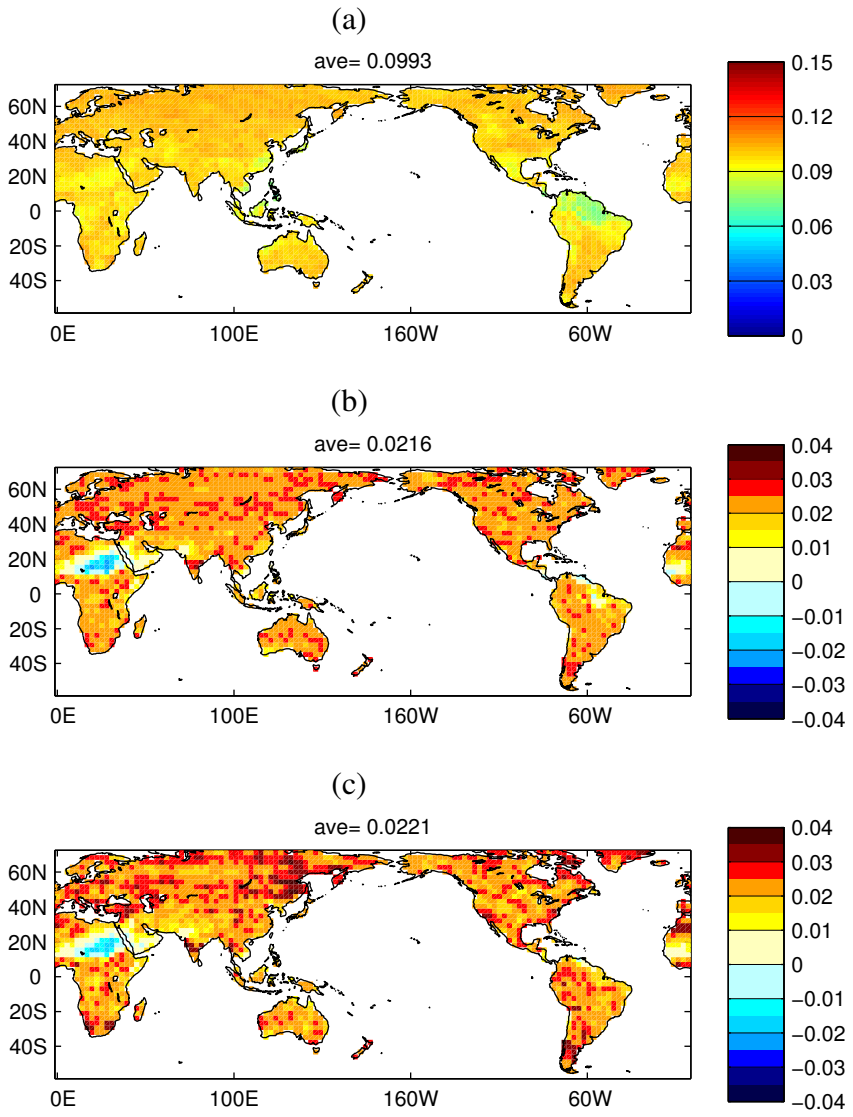



Figure 8. (a) RMS error of the counting estimate of the below-normal tercile probability with ensemble size 20. The RMS error of the counting error minus that of the (b) Gaussian fit and (c) the GLM based on the ensemble mean. The gridpoint averages are shown in the titles.

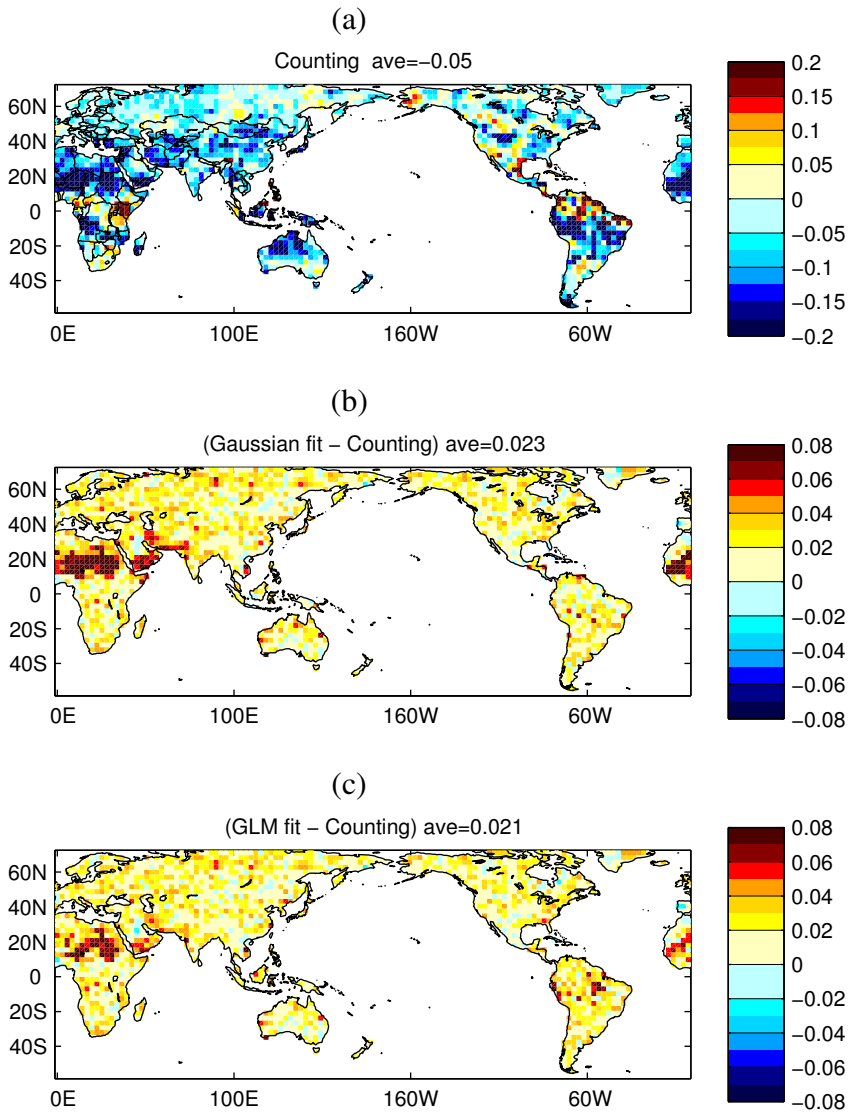



Figure 9. RPSS of (a) the counting-based probabilities and its difference with that of the (b) Gaussian and (c) GLM estimated probabilities. Positive values in (b) and (c) correspond to increased RPSS compared to counting. The gridpoint averages are shown in the titles.

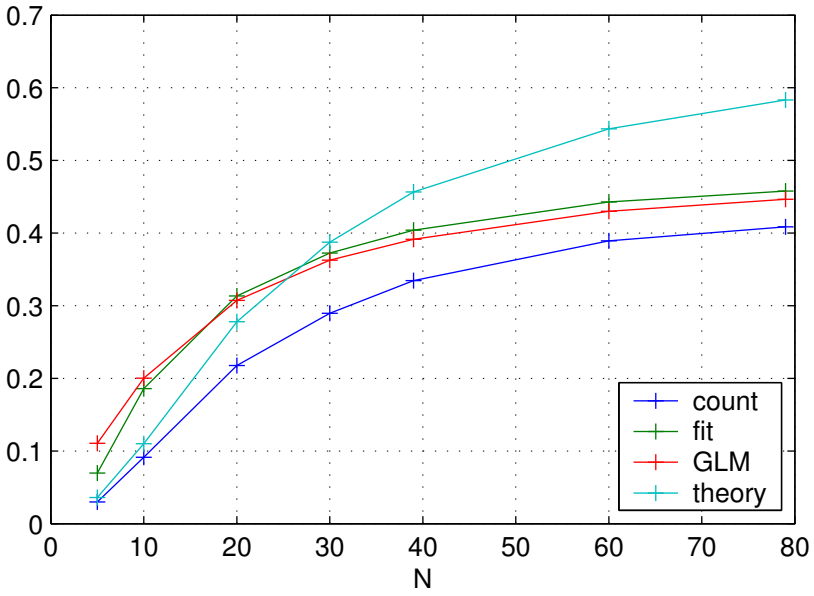



Figure 10. The fraction of land points with RPSS$> 0$.

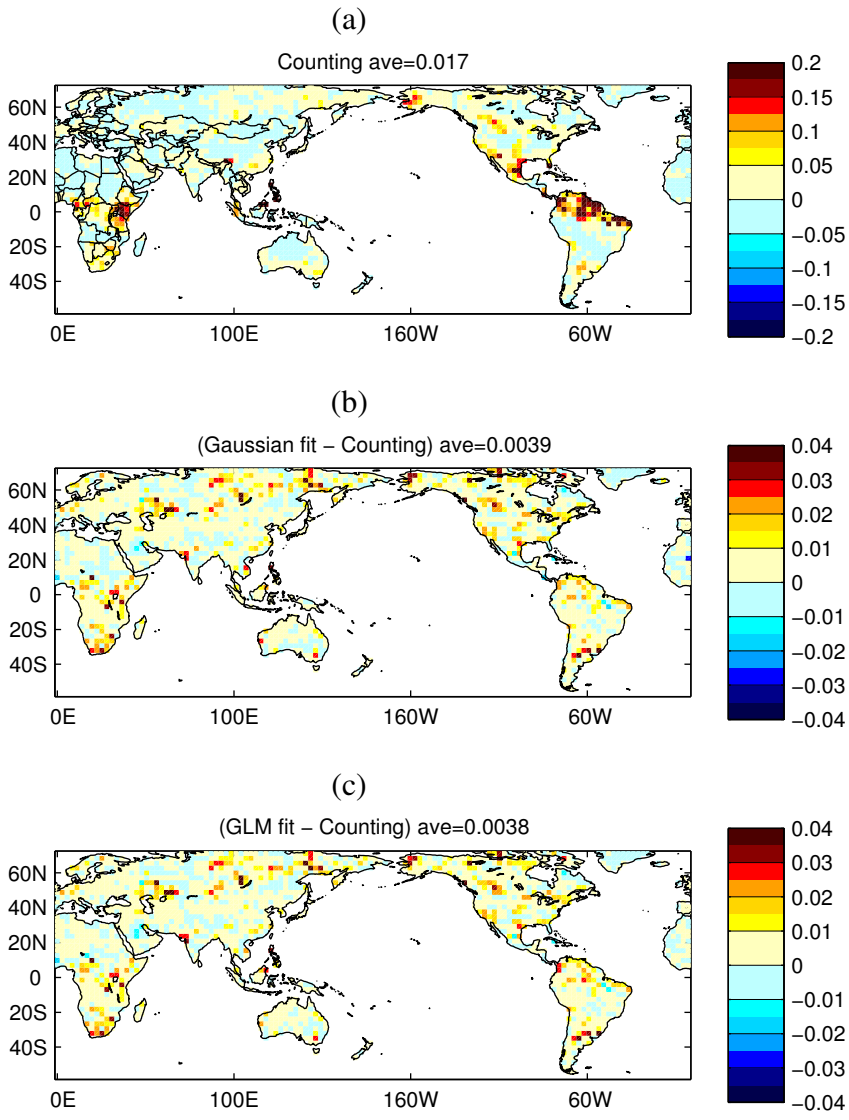



Figure 11. As in Fig. 9 but for the Bayesian calibrated probabilities.

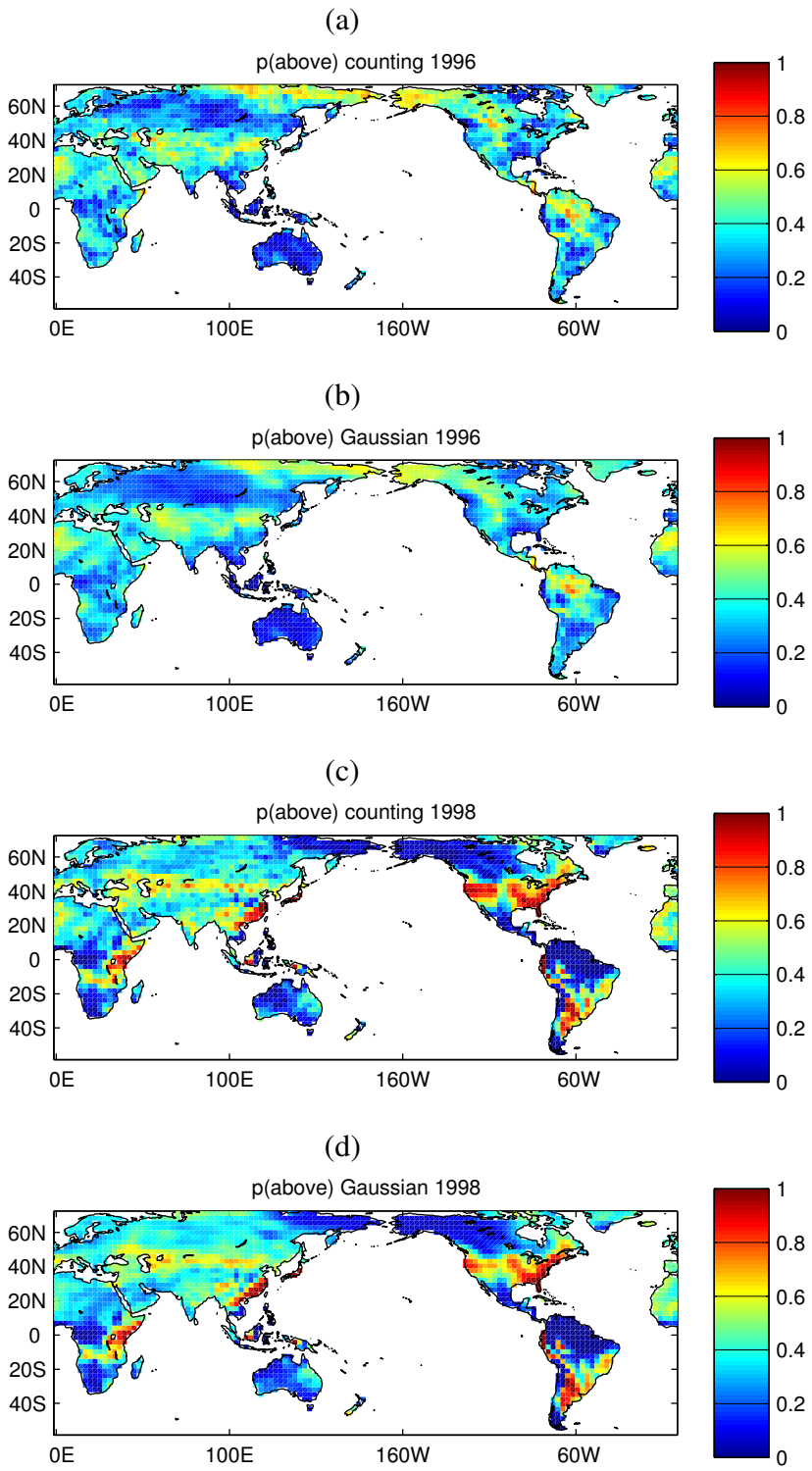



Figure 12. Probability of above-normal precipitation for DJF 1996 estimated by (a) counting and (b) Gaussian fit, and DJF 1998 using (c) counting and (d) Gaussian fit.