Estimation of seasonal precipitation tercile-based catagorical probabilities from ensembles

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

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

November 21, 2005

ABSTRACT

A simple method of estimating precipitation tercile-based categorical probabilities from seasonal ensemble forecasts is to count the number of ensemble members that fall in each tercile-based category. Another method is to estimate categorical probabilities from fitted parametric distributions. Here we investigate how sampling error due to finite ensemble size effects the counting and parametric methods as well as an empirical approach where categorical probabilities are obtained from a generalized linear model. The methods are compared in an idealized setting using analytical results showing the scaling of sampling error with method, ensemble size, and level of predictability. The robustness of the theoretical results is explored in a more practical setting with seasonal precipitation simulated by a general circulation model. In addition to determining the relative accuracies of the different methods, the analysis quantifies the relative importance of the ensemble mean and spread in estimating tercile probabilities.

1. Introduction

Seasonal climate forecasts are necessarily probabilistic, so that forecast information is most completely characterized by a probability density function (pdf). Differences between the climatology pdf and the forecast pdf represent predictability, and several measures of this difference have been developed to quantify useful predictability in seasonal forecasts and simulations (Kleeman 2002; DelSole 2004; Tippett et al. 2004). Quantile probabilities provide a coarse-scale indication of the difference between the forecast and climatological pdf appropriate for ensemble methods (Kleeman and Majda 2005). The International Research Institute for Climate and Society (IRI) seasonal forecasts of precipitation and temperature are 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). The accurate estimate of quantile probabilities is important both for quantifying seasonal predictability and for making climate forecasts.

In single-tier forecasts, initial conditions of the ocean-atmosphere system are the source of predictability, and ensembles of coupled model forecasts give samples of the atmosphere-ocean system consistent with the initial condition and its uncertainty. In simulations and two-tier seasonal forecasts, the goal is to know the range and likelihood of possible atmospheric states consistent with a particular sea surface temperature (SST). In simulation mode, forcing an ensemble of an atmospheric general circulation model (GCM) with observed SST gives a sample of equally likely atmospheric responses to SST forcing. Tercile probabilities must be estimated from finite ensembles in either setting. Tercile probabilities can be estimated from the ensemble in two different ways. A simple nonparametric estimate of the tercile probabilities is a count of the number of ensemble members in each category. Alternatively, the tercile probabilities can be estimated parametrically by fitting the ensemble by a pdf with some adjustable parameters. The counting method has the advantage of making no assumptions about the forecast pdf. Both approaches are affected by sampling error caused by a finite ensemble size, though to different degrees. This paper is about the relative merits of these two approaches for seasonal simulation and forecasting of precipitation tercile probabilities. We choose to 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 which have been previously examined.

Richardson (2001) examined the effect of ensemble size on the Brier skill score, reliability and economic value determined from a simple cost-loss decision model when probabilities are obtained by counting. The finite size of the ensemble had an adverse effect on the Brier skill score with low-skill regions begin more negatively affected by small ensemble size. The effect of finite ensemble size on reliability was, on average, overconfidence in the forecast probabilities. Furthermore, changes in ensemble size that cause only modest changes in Brier skill score were found to lead to large changes in economic value, particularly for extreme events.

Several authors have compared estimation of quantile probabilities by counting and pdf fitting. Wilks (2002) fit numerical weather prediction ensembles with a Gaussian or a mixture of Gaussians and showed that fitted distributions give more accurate estimations of quantile values, especially for quantiles near the extremes of the distribution. Kharin and Zwiers (2003) showed using Monte Carlo simulations that when the forecast variable is Gaussian, using a Gaussian estimator was more accurate than counting. They showed with hindcast data that the Brier skill score of hindcasts of 700 mb temperature and 500 mb height was improved when probabilities were estimated from a Gaussian with constant variance as compared with counting; fitting a Gaussian with time-varying variance gave inferior results. Hamill et al. (2004) used a generalized linear model (GLM; logistic regression) to relate tercile probabilities with the ensemble mean in the context of 6-10 day and week 2 surface temperature and precipitation forecasts. The ensemble spread was not found to be a useful predictor of tercile probabilities.

Here we present some analytical results for the effects of ensemble size and level of predictability on tercile probability for counting and Gaussian estimates. These analytical results permit comparison of the error levels of the counting and Gaussian estimates and their dependence on ensemble size and predictability. The calculation of the analytical results uses some simplifying assumptions whose importance we assess by comparing the analytical estimates with empirical estimates obtained by sub-sampling from a 79-member ensemble of GCM simulations. Ensemble estimated tercile probabilities form the inputs to IRI's forecast system that includes calibration (Barnston et al. 2003; Robertson et al. 2004), and we expect that reducing sampling error will improve forecast skill. Reduction of sampling error should also benefit the estimation of the calibration parameters. However, model error and the calibration process may confound the effect of reducing sampling error. Therefore we first look at sampling error in a perfect model setting by comparing the tercile probabilities of a large ensemble to those estimated from smaller ensembles. Then we look at the effect of using probabilities having smaller sampling error on the skill of the simulations with and without calibration. In the perfect model setting, we investigate whether the time-varying ensemble variance is a useful factor in estimating the tercile probabilities. We use a generalized linear model (GLM) to infer tercile probabilities from ensemble statistics. We show that the GLM used is equivalent to fitting a Gaussian model for Gaussian variables. The GLM approach allows us to identify which estimated ensemble statistics, in particular mean and variance, are robustly related to tercile probability and hence to predictability.

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 model error. A summary and conclusions are given in section 5.

2. Data

Model simulated precipitation data comes from a 79-member ensemble of T42 ECHAM 4.5 GCM (Roeckner et al. 1996) simulations forced with observed SST for the period Dec 1950 to Feb 2003. We use seasonal averages of the three month period of 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 are use unprocessed model simulated precipitation, many of the calculations were repeated using Box-Cox transformed data; the Box-Cox transformation

$$x_{BC} = \lambda^{-1} \left(x^{\lambda} - 1 \right) \, , \tag{1}$$

makes the data approximately Gaussian and depends on the parameter λ; $x_{BC} = \log x$ for $\lambda = 0$. Positive skewness is the usual non-Gaussian aspect for precipitation. The best value of λ is found by maximizing the log likelihood function. Figure 1 shows the geographical distribution of the values of λ and an indication of the deviation of the data from Gaussianity; we only allow a few values of λ, 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.

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

$P_p(n|N)$ gives the probability of there being exactly n members in the category; p is the true tercile probability. The expected number of members in the tercile category is

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

where the notation $\langle \cdot \rangle$ denotes expectation. Consequently, the expected value of the counting estimate p_N is the true probability p. However, having a limited ensemble size generally causes any single realization of p_N not to exactly equal p. The variance of the counting estimate p_N is

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

where we have used the fact that the variance of the binomial frequency 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, and Fig. 2a 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 $2/9N$, and the percentiles are obtained by inverting the cumulative distribution function of the sample error. Figure 2a 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 variance of the counting estimate depends on the true probability p and vanishes as p approaches either 0 or 1. Figures 2b and 2c shows curves of the 16th and 84th percentiles and of the mean plus and minus the standard deviation as functions of p for different values of N.

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 these probability values differ from the climatological value of $p = 1/3$ is an indication of predictability, with larger deviations indicating more predictability. Intuitively, we expect regions with more predictability to suffer less from sampling error on average, since enhanced predictability implies more reproducibility among ensemble members. In fact, when the changes in the probability p are due to changes in the mean of a Gaussian distribution with constant variance, the variance of the counting estimate (see the Appendix for details) is approximately

$$\left\langle (p - p_N)^2 \right\rangle \approx -\frac{0.002856}{N} + \frac{0.225079}{N\sqrt{1+S^2}} \approx \frac{2}{9N\sqrt{1+S^2}}\,, \tag{4}$$

where the signal-to-noise ratio S is the ratio of the ensemble mean standard deviation to the mean of the ensemble spread standard deviation; the correlation level associated with S is $S/\sqrt{1+S^2}$. In the present context, predictability is associated with both smaller ensemble spread relative to climatology and smaller average variance of the counting estimate. The relation in (4) has the practical application 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, though 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* of the tercile probabilities is found by fitting the 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 underlying distribution that the ensemble distribution represents and (ii) sampling error in the mean and variance estimates 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 ensembles. For ensembles with known climatological variance and no predictability ($S = 0$), the dependence of the Gaussian fit tercile probability variance on ensemble size derived in the appendix is given approximately by

$$\left\langle (p - g_N)^2 \right\rangle \approx \frac{e^{-x_b^2}}{2\pi N} \approx \frac{0.1322}{N}\,; \tag{5}$$

the right tercile boundary x_b is given by $x_0 = \Phi^{-1}(2/3) \approx 0.4307$ where Φ is the normal cumulative distribution function; the approximation becomes more accurate as the ensemble size N increases. The variance increases by the factor $(N-1)/(N-3)$ if the variance of the Gaussian is not constant and must be also be estimated from the ensemble of size N. Comparing (3) and (5), 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 actually Gaussian with known climatological variance and no predictability ($S = 0$). 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 (5) 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 for estimation of the median, $x_0 = 0$ and the the variance of the Gaussian estimate is about 28% smaller than that of the counting estimate; in the case of deciles, $x_0 = 1.2816$, and the variance of the Gaussian estimated decile probability is about 66% smaller than that of the counting estimate. The approximation in (5) is only accurate for higher quantiles when the ensemble size is sufficiently large.

If the signal-to-noise ratio S is not zero, the variance of the Gaussian fit tercile probability derived in the appendix is approximately

$$\frac{e^{-\frac{1+S^2}{1+2S^2}x_0^2}}{2\pi N\sqrt{1+2S^2}} \,. \tag{6}$$

To compare this value with the counting estimate variance in (4), Fig. 3 shows the ratio of the standard deviations of the counting and the Gaussian fit estimates as a function of the signal-to-noise ratio S. The Gaussian fit estimate has smaller variance for all values of S, with its advantage over the counting estimate increasing slightly as the signal-to-noise ratio increases to levels exceeding unity, as occurs mainly in the tropics.

c. Estimates from generalized linear models

Generalized linear models (GLMs) offer a parametric estimate of quantile probabilities without the 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

$$R = \sum_i a_i y_i \,, \tag{7}$$

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

$$g(p) = R = \sum_i a_i y_i \,. \tag{8}$$

The parameters a_i are then found by maximum likelihood estimation. Here, the GLMs are developed with the standardized mean and standard deviation as explanatory variables and p given by the counting estimate.

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, that is, we define

$$g(p) \equiv \sqrt{2}\,\mathrm{erf}^{-1}(1-2p)\,. \tag{9}$$

Results using the logit function 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 standardized ensemble mean and standard deviation. When the ensemble distribution is Gaussian with constant variance, $g(p)$ is linearly related to the ensemble mean and this assumption is exactly satisfied. To see this, suppose that the ensemble has mean f and variance σ_e. Then the probability p of the below normal category is

$$p = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-x_b} e^{-\frac{(x-f)^2}{2\sigma_e^2}}\, dx = \frac{1}{2}\left(1 - \mathrm{erf}\left(\frac{-x_b - f}{\sqrt{2}\sigma_f}\right)\right)\,, \tag{10}$$

where x_b is the tercile of the climatological distribution. Then,

$$g(p) = \sqrt{2}\,\mathrm{erf}^{-1}\left(\mathrm{erf}\left(\frac{-x_b - f}{\sqrt{2}\sigma_f}\right)\right) = \frac{-x_b - f}{\sigma_f}\,. \tag{11}$$

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 estimates of the mean.

We show an example with synthetic data to give an indication of the robustness of the GLM estimate when the population that ensemble represents does not have a Gaussian distribution. We take the ensemble pdf to be a gamma distribution with parameters (2,1). The pdf is asymmetric and has a positive skew (see Fig. 4a). 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. 4b). 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 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

$$RPS \equiv \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{3}(F_{i,j}-O_{i,j})^2\,, \tag{12}$$

where M is the number of forecasts, $F_{i,j}$ ($O_{i,j}$) is the cumulative distribution function of the ith forecast (observation) at the jth category. The observation pdf 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 ith forecast, and that the terms with $j = 3$ 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. Note that $\mathcal{O}_B$, $\mathcal{O}_N$, and $\mathcal{O}_A$ collectively represent the uncertainty of the observation, 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, but can be estimated making the perfect model assumption. The expected (with respect to the observations) RPS of a particular forecast is found by summing the RPS for each possible category of the observation multiplied by its likelihood, giving

$$\begin{aligned}\langle RPS\rangle &= \mathcal{O}_B\left[(P_B-1)^2+(P_B+P_N-1)^2\right]+\mathcal{O}_N\left[P_B^2+(P_B+P_N-1)^2\right]+\mathcal{O}_A\left[P_B^2+(P_B+P_N)^2\right]\\ &= \mathcal{O}_B\left[(P_B-1)^2+P_A^2\right]+\mathcal{O}_N\left[P_B^2+P_A^2\right]+\mathcal{O}_A\left[P_B^2+(1-P_A)^2\right]\;.\end{aligned} \tag{13}$$

The observation and forecast probabilities are equal for a perfect forecast model, and using (13), the expected RPS (denoted RPS_{perfect}) of a perfect forecast model is

$$\begin{aligned}RPS_{\text{perfect}} &\equiv \mathcal{O}_B\left[(\mathcal{O}_B-1)^2+\mathcal{O}_A^2\right]+\mathcal{O}_N\left[\mathcal{O}_B^2+\mathcal{O}_A^2\right]+\mathcal{O}_A\left[\mathcal{O}_B^2+(1-\mathcal{O}_A)^2\right]\\ &= \mathcal{O}_B(1-\mathcal{O}_B)+\mathcal{O}_A(1-\mathcal{O}_A)\,.\end{aligned} \tag{14}$$

Note that the expected RPS of a perfect forecast model is not zero unless the probability of a category is one, or zero. Figure 5a shows the time-averaged value of RPS_{perfect} for the 79 member ECHAM 4.5 GCM simulated precipitation data using observed SST forcing for DJF 1951-2003. This is a perfect model measure of skill with small values of RPS_{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 skills are close to the no-skill limit of 4/9 in much of the extratropics.

Suppose that the forecast and observation probabilities are different, with $P_B = \mathcal{O}_B + \epsilon_B$ and $P_A = \mathcal{O}_A + \epsilon_A$, and that ϵ_B and ϵ_A represent sampling error; that is, the forecast model is perfect, but the ensemble size is finite. If the forecast probabilities are unbiased and $\langle\epsilon_B\rangle = \langle\epsilon_A\rangle = 0$, then averaging over realizations of the sampling error gives

$$\begin{aligned}\langle RPS\rangle &= RPS_{\text{perfect}} + \mathcal{O}_B\left\langle \epsilon_B^2+\epsilon_A^2\right\rangle + \mathcal{O}_N\left\langle \epsilon_B^2+\epsilon_A^2\right\rangle + \mathcal{O}_A\left\langle \epsilon_B^2+\epsilon_A^2\right\rangle \\ &= RPS_{\text{perfect}} + \left\langle \epsilon_B^2+\epsilon_A^2\right\rangle \,.\end{aligned} \tag{15}$$

This means that the previous results for the variance of the tercile probability estimates can be directly used to quantify how sampling error increases RPS. In particular, if the sampling error is associated with the counting estimate whose variance is given by (3), then

$$\left\langle \epsilon_B^2+\epsilon_A^2\right\rangle = \frac{1}{N}\left(\mathcal{O}_B(1-\mathcal{O}_B)+\mathcal{O}_A(1-\mathcal{O}_A)\right)\,, \tag{16}$$

and

$$\langle RPS\rangle = \left(1+\frac{1}{N}\right) RPS_{\text{perfect}}\,. \tag{17}$$

The RPSS is defined using the RPS and a reference forecast defined to have zero skill, such as climatology:

$$RPSS = 1 - \frac{RPS}{RPS_{\text{ref}}}\,, \tag{18}$$

where RPS_{ref} is the RPS of the reference forecast. The expected RPS of a climatological forecast with $P_B = P_N = P_A = 1/3$ is

$$RPS_{\text{clim}} = \frac{2}{9} + \frac{1}{3}\left(\mathcal{O}_B + \mathcal{O}_A\right)\,. \tag{19}$$

and

$$\begin{aligned}RPSS &= 1 - \left(1+\frac{1}{N}\right)\frac{RPS_{\text{perfect}}}{RPS_{\text{clim}}}\,, \\ &= \frac{(N+1)RPSS_{\text{perfect}} - 1}{N}\,,\end{aligned} \tag{20}$$

where $RPSS_{\text{perfect}} \equiv 1 - RPS_{\text{perfect}}/RPS_{\text{clim}}$. Figure 5b shows the time-averaged value of $RPSS_{\text{perfect}}$ for the GCM simulated precipitation data. The relation between RPSS and ensemble size is the same as that for the Brier skill score (Richardson 2001). If the error variance of the estimate differs from that of the counting estimate by the factor α, as for example if the Gaussian fit estimate is used, then

$$RPSS = \frac{(N+\alpha)RPSS_{\text{perfect}} - \alpha}{N}\,, \tag{21}$$

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 expression in (4) for the average variance of the counting estimate was derived assuming that the ensemble mean and its uncertainty are both Gaussian distributions, and therefore, depends on the signal-to-noise ratio. We now compare it with the average variance of the counting estimate computed by sub-sampling from a large ensemble of GCM simulations to see how well the simple expression in (4) described the behavior of real data. We compute the variance of the counting estimate of tercile probability by selecting two independent samples of size N (without replacement) from the ensemble of GCM simulations and using them to compute two values of the counting estimate probabilities denoted p_N and p'_N; the ensemble size of 79 and requirement of independence limits the maximum value of N to 39. Then the variance of the difference between the two counting estimates p_N and p'_N is twice the error variance of the counting estimate since

$$\left\langle (p_N - p'_N)^2 \right\rangle = \left\langle ((p_N - p) + (p - p'_N))^2 \right\rangle = 2\left\langle (p - p_N)^2 \right\rangle\,, \tag{22}$$

where we use the fact that the sampling errors $(p_N - p)$ and $(p - p'_N)$ are uncorrelated. The averages in (22) are with respect to time and realizations (1000) of the two independent samples.

We expect close agreement between the sub-sampling results and the analytical results of (4) in regions where there is little predictability and the signal-to-noise ratio S is small, since for $S = 0$ the analytical result is exact. In regions where $S > 0$ and some years there are shifts from equal odds, we expect that the average counting variance still decreases as $1/N$ but there is no guarantee that the analytical result obtained using a Gaussian approximation will be a useful description of actual model precipitation. However, Fig. 6 shows that in the land gridpoint average, the variance of the counting estimate is very well described by the analytical result in (4). The theoretical value slightly exceeds the empirical value; one reason for this difference is that the leading order term neglected in the approximation is negative. The empirical value of the average variance of the below-normal tercile probability is slightly less than that of the above-normal tercile probability.

Figure 7a shows the spatial variation of the factor $\sqrt{2}/3(1+S^2)^{1/4}$ appearing in (4). This factor can be interpreted as the standard deviation 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 factor can also be obtained empirically from sub-samples of varying size, and the difference between the theoretical factor and the empirical estimate is mostly on the order of a few percent (see Figs. 7b and 7c); the values are mostly negative, likely for the reason mentioned above concerning the sign of the leading-order neglected term. The factor for the below-normal tercile is generally smaller than that for the above-normal tercile, consistent with the gridpoint-averaged results. The agreement between the theoretical and the empirical factor is slightly better for the Box-Cox transformed data (not shown).

We now use sub-sampling of the GCM simulated precipitation data to compare the three estimation methods discussed in the previous section: counting, Gaussian fit and GLM. The method used above where the estimator is applied to two independent samples is not an appropriate way of computing the error variance of each estimator because the Gaussian fit and GLM estimators may be biased. For instance, one could imagine that the difference of the Gaussian fit estimator applied to two independent samples is small but the error of the estimate is not. 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; the 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 and the 40-member counting estimate. The variance of the difference can be decomposed into error variance contributions from q_N and p_{40}:

$$\begin{aligned}\langle (q_N - p_{40})^2 \rangle &= \langle (q_N - p + p - p_{40})^2 \rangle \\ &= \langle (q_N - p)^2 \rangle + \langle (p - p_{40})^2 \rangle \\ &\approx \langle (q_N - p)^2 \rangle - \frac{0.002856}{40} + \frac{0.225079}{40\sqrt{1+S^2}}\,,\end{aligned} \tag{23}$$

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

$$\langle (q_N - p)^2 \rangle \approx \langle (q_N - p_{40})^2 \rangle + \frac{0.002856}{40} - \frac{0.225079}{40\sqrt{1+S^2}}\,. \tag{24}$$

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 have 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 8a 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. 8a) and is larger than that of the parametric estimates. The one parameter GLM and constant variance Gaussian fit have similar size 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 two parameter estimates (Fig. 8b).

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 do not improve estimates of the tercile probabilities using the range of ensemble sizes considered here (Kharin and Zwiers 2003).

Figure 9 shows the spatial features of the rms error of the below-normal tercile probability estimates for ensemble size 20. We see that using a Gaussian with constant variance or a GLM based only 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 time-varying and when both the mean and standard deviation are used in the GLM are similar to those in Fig. 9, but the overall error levels are slightly higher,

b. RPSS

Having evaluated the three probability estimation methods in the perfect model setting where we have asked how closely they match the probabilities from an infinite size ensemble, we now use the RPSS to compare the 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 10 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 11 shows the sum over land gridpoints of the positive values of RPSS and 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 gridpoints with positive RPSS and a higher "total" 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 (20) 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 (20) for $RPSS_{\text{perfect}}$; we expect that in this case sampling error dominates model error and the relation in (20) holds approximately. Then we use (20) to compute the gridpoint averaged RPSS for other values of N; these values are shown in the theory curve in Fig. 11. In the absence of model error, the count and theory curves of RPSS in Fig. 11 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 (20).

The presence of model error means that some calibration of the model output with observations is needed. 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) where the weights given to the GCM probability relative to the climatology probability ($p = 1/3$) are chosen to maximize the likelihood of the observations. There is cross-validation in the sense that the weights are computed with one sample of size N and the RPSS is computed by applying those weight to different sample of size N and then comparing the result with observations. The calibrated counting-estimated probabilities still have slightly negative RPSS in some areas (Fig. 12a) but the overall amount of positive RPSS is increased compared to the uncalibrated simulations (compare with Fig. 10a); 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. 12b,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 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 if fitting leads to overly conservative probabilities. Figure 13 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 chosen ensemble size and probability estimation technique are related to the accuracy of tercile probability estimates. 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 sampling variance of the counting estimate is a function of the probability and the ensemble size. For Gaussian data, the average variance was shown to depend on the ensemble size and the signal-to-noise ratio. The Gaussian fit estimate assumes that the ensemble can be described by a Gaussian distribution, and when this is true, we have shown analytically that this estimate has an approximately 23% smaller standard deviation of error than the counting estimate for tercile probabilities. The advantage of the Gaussian fit over the counting estimate is equivalent to fairly substantial increases in ensemble size. 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 variance of the tercile probability estimates is important because it is closely related to the degree to which the ranked probability skill score (RPSS) is degraded by sampling error.

Some of the theoretical results are obtained assuming that the ensemble is well-described by a Gaussian. We test their robustness 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 to understand how sampling error depends on ensemble size and level of predictability. Although the GCM simulated precipitation is not Gaussian, the Gaussian fit estimate had smaller error than the counting estimate. The behavior of the GLM estimate was similar to the Gaussian fit estimate. The 1-parameter (ensemble mean) parametric estimators had the best performance; adding ensemble variance as a parameter did not reduce error. This suggests that with the moderate ensemble sizes typically used, the varying ensemble spreads from forecasts to forecast are largely dominated by sampling variability.

The reduced sampling error of the Gaussian fit and GLM translates into better simulation skill when the tercile probabilities are compared to actual observations. We compared the dependence of the RPSS on ensemble size under the perfect model assumption and with actual observations. Although RPSS increases with ensemble size, model error reduces the rate of improvement compared to the perfect model case. Calibration improves RPSS, regardless of the probability estimator used. However, estimators with larger sampling error retained 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 without damping the strength of the tercile probabilities. The Gaussian fit probabilities are spatially relatively smoother than those estimated by counting.

In summary, our main conclusion is that carefully applied parametric estimators provide more accurate tercile probabilities than do counting estimates. This conclusion is completely rigorous for ensembles 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. IRI is supported by its sponsors and NOAA Office of Global Programs Grant number NA07GP0213.

APPENDIX

Error in estimating tercile probabilities

a. Derivation of the counting variance

Suppose we estimate the probability of a tercile category, for instance, the below-normal category, by drawing N times from the forecast pdf and counting the number of times that the draw falls in the below-normal category. The binomial distribution $P_p(n|N)$ gives the probability of obtaining exactly n draws in the below-normal category where p is the population probability of the below-normal category. The variance of the counting estimate of p is

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

where we have used the facts that the mean and variance of the binomial distribution are pN and $N(1-p)p$, respectively. If the tercile probability p is constant, say $p = 1/3$ as in regions where there is no skill, then the variance is $2/(9N)$. In regions where there is no skill, the variance depends only on the sample size and the number of climatologically equi-probable categories.

Generally, the tercile probability p depends on the forecast and climatology distributions and is a random variable. Therefore the variance is $(\langle p \rangle - \langle p^2 \rangle)/N$ where $\langle \cdot \rangle$ denotes expectation. Now we compute the average variance for a quantity whose distribution is Gaussian with a mean that varies. Suppose that the uncertainty of the expected forecast f is normally distributed with zero mean and variance σ_e^2. Then the probability p of the below-normal category is the integral of the forecast pdf up to the left tercile boundary $-x_b$:

$$p = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-x_b} e^{-\frac{(x-f)^2}{2\sigma_e^2}} \, dx = \frac{1}{2}\left(1 - \mathrm{erf}\left(\frac{-x_b - f}{\sqrt{2}\sigma_e}\right)\right)\,, \tag{A.2}$$

where erf denotes the error function. If we further assume that the expected forecast f is a normally distributed random variable with zero mean and variance σ_f^2, the tercile boundary x_b of the climatological pdf is $x_0\sqrt{\sigma_e^2+\sigma_f^2}$ where $x_0 \approx 0.430727$.

With these assumptions we can compute the terms in the variance $(\langle p\rangle - \langle p^2\rangle)$. First,

$$\langle p\rangle = \frac{1}{2\pi\sigma_e\sigma_f}\int_{-\infty}^{\infty}\int_{-\infty}^{-x_b} e^{-\frac{(x-f)^2}{2\sigma_e^2}} e^{-\frac{-f^2}{2\sigma_f^2}}\ dx\ df = \frac{1}{3}\,. \tag{A.3}$$

Then,

$$\langle p^2\rangle = \frac{1}{(2\pi)^{3/2}\sigma_e^2\sigma_f}\int_{-\infty}^{\infty}\int_{-\infty}^{-x_b}\int_{-\infty}^{-x_b} e^{-\frac{(x-f)^2}{2\sigma_e^2}} e^{\frac{(y-f)^2}{2\sigma_e^2}} e^{-\frac{f^2}{2\sigma_f^2}}\ dx\ dy\ df\,. \tag{A.4}$$

Performing two of the integrals gives:

$$\langle p^2\rangle = \frac{g}{2\sqrt{2\pi}}\int_{-\infty}^{-x_b}\left(1-\mathrm{erf}\left(\frac{x(1-g^2)+x_b}{\sqrt{2}\sqrt{2-g^2}}\right)\right)e^{-\frac{g^2x^2}{2}}\ dx\,, \tag{A.5}$$

where $g \equiv (1+\sigma_f^2/\sigma_e^2)^{-1/2}$. Changing variables so that $x \longrightarrow x/g$. give

$$\langle p^2\rangle = \frac{1}{2\sqrt{2\pi}}\int_{-\infty}^{-x_0}\left(1-\mathrm{erf}\left(\frac{x/g(1-g^2)+x_0/g}{\sqrt{2}\sqrt{2-g^2}}\right)\right)e^{-\frac{x^2}{2}}\ dx\,. \tag{A.6}$$

We perform a series expansion about $g=1$ (signal-to-noise ratio $S=0$) and find

$$\langle p^2\rangle_{g=1} = 1/9\,, \tag{A.7}$$

and

$$\frac{d}{dg}\langle p^2\rangle_{g=1} = -0.225079\,. \tag{A.8}$$

This gives that

$$\langle p^2\rangle = 1/9 - 0.225079(g-1) + \mathcal{O}(g-1)^2 = 0.336190 - 0.225079g + \mathcal{O}(g-1)^2\,. \tag{A.9}$$

Finally, the average variance is approximately

$$\frac{\langle p\rangle - \langle p^2\rangle}{N} \approx -\frac{0.002856}{N} + \frac{0.225079}{N\sqrt{1+S^2}} \approx \frac{2}{9N\sqrt{1+\sigma_f^2/\sigma_e^2}}\,. \tag{A.10}$$

b. Error in estimating tercile probabilities by Gaussian fitting

Suppose we fit a Gaussian to the ensemble X_i, finding its mean f and sample standard deviation s_e defined by

$$s_e \equiv \frac{1}{n-1} \sum_{i=1} (X_i - f)^2 \,. \tag{A.11}$$

Based on this information and using (A.2), the probability of the below category is

$$\frac{1}{2}\left(1 - \operatorname{erf}\left(\frac{-x_b - f}{\sqrt{2} s_e}\right)\right) \,, \tag{A.12}$$

where $x_b = x_0/\sigma_c$ and σ_c is the climatological variance.

Suppose that there is no forecast signal, i.e., the expected mean of the ensemble is zero, the true tercile probability is 1/3 and $s_e = \sigma_c$. Then, the variance of the Gaussian fit estimate is

$$\left(\frac{1}{3} - \frac{1}{2}\left(1 - \operatorname{erf}\left(\frac{-x_b - f}{\sqrt{2} s_e}\right)\right)\right)^2 = \frac{f^2 e^{-x_0^2}}{2 s_e^2 \pi} + \mathcal{O}(f^3) \,, \tag{A.13}$$

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

$$\left\langle \frac{f^2 e^{-x_0^2}}{2 s_e^2 \pi} \right\rangle = \frac{e^{-x_0^2}}{2\pi N} \frac{N-1}{N-3} \approx \frac{0.1322}{N} \frac{N-1}{N-3} \,. \tag{A.14}$$

If the forecast variance s_e is known then the expected squared error of the tercile probability is

$$\left\langle \frac{f^2 e^{-x_0^2}}{2 \sigma_c^2 \pi} \right\rangle = \frac{e^{-x_0^2}}{2\pi N} \approx \frac{0.1322}{N} \,, \tag{A.15}$$

since $\langle f^2 \rangle = \sigma_c^2/N$.

Suppose that the signal-to-noise ratio S is not zero, and the forecast variance s_e is constant and known. The variance is

$$\left\{\frac{1}{2}\left(1 - \operatorname{erf}\left(\frac{-x_b - f}{\sqrt{2} s_e}\right)\right) - \frac{1}{2}\left(1 - \operatorname{erf}\left(\frac{-x_b - F}{\sqrt{2} s_e}\right)\right)\right\}^2 \,, \tag{A.16}$$

where F is the true mean and f is the mean estimated from the N-member ensemble. Expanding this expression in a Taylor series in in powers of $(F - f)$ about $F = f$ to get that the variance is

$$\frac{(f - F)^2 e^{-(F/s_f + \sqrt{1+S^2}x_0)^2}}{2\pi s_f^2} + \mathcal{O}(F - f)^3 \,. \tag{A.17}$$

Keeping only the leading order terms and using the fact that the quantity $(F - f)/s_f$ has a t-distribution, leads to the expression

$$\frac{e^{-(F/s_f + \sqrt{1+S^2}x_0)^2}}{2\pi N} \,, \tag{A.18}$$

for variance averaged with respect to realizations of the ensemble. Then we use the fact that the true mean F has a Gaussian distribution to average over forecasts and obtain that the average variance is

$$\frac{e^{-\frac{1+S^2}{1+2S^2}x_0^2}}{2\pi N\sqrt{1 + 2S^2}} \,. \tag{A.19}$$

REFERENCES

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.

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

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. Majda, 2005: Predictability in a model of geophysical turbulence. *J. Atmos. Sci.*, **62**, 2864–2879.

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

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.

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.

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

List of Figures

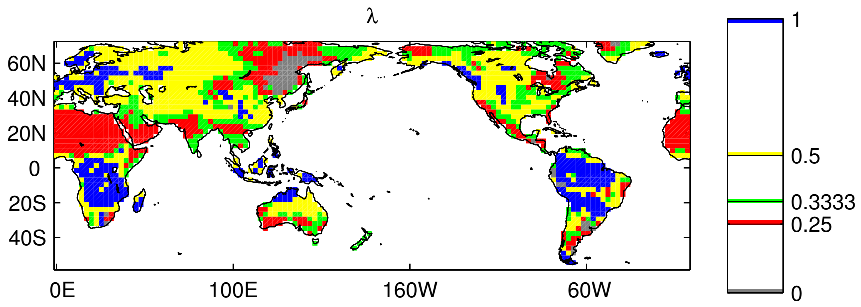


Figure 1. Spatial distribution of λ appearing the Box-Cox transformation of Eq. 1.

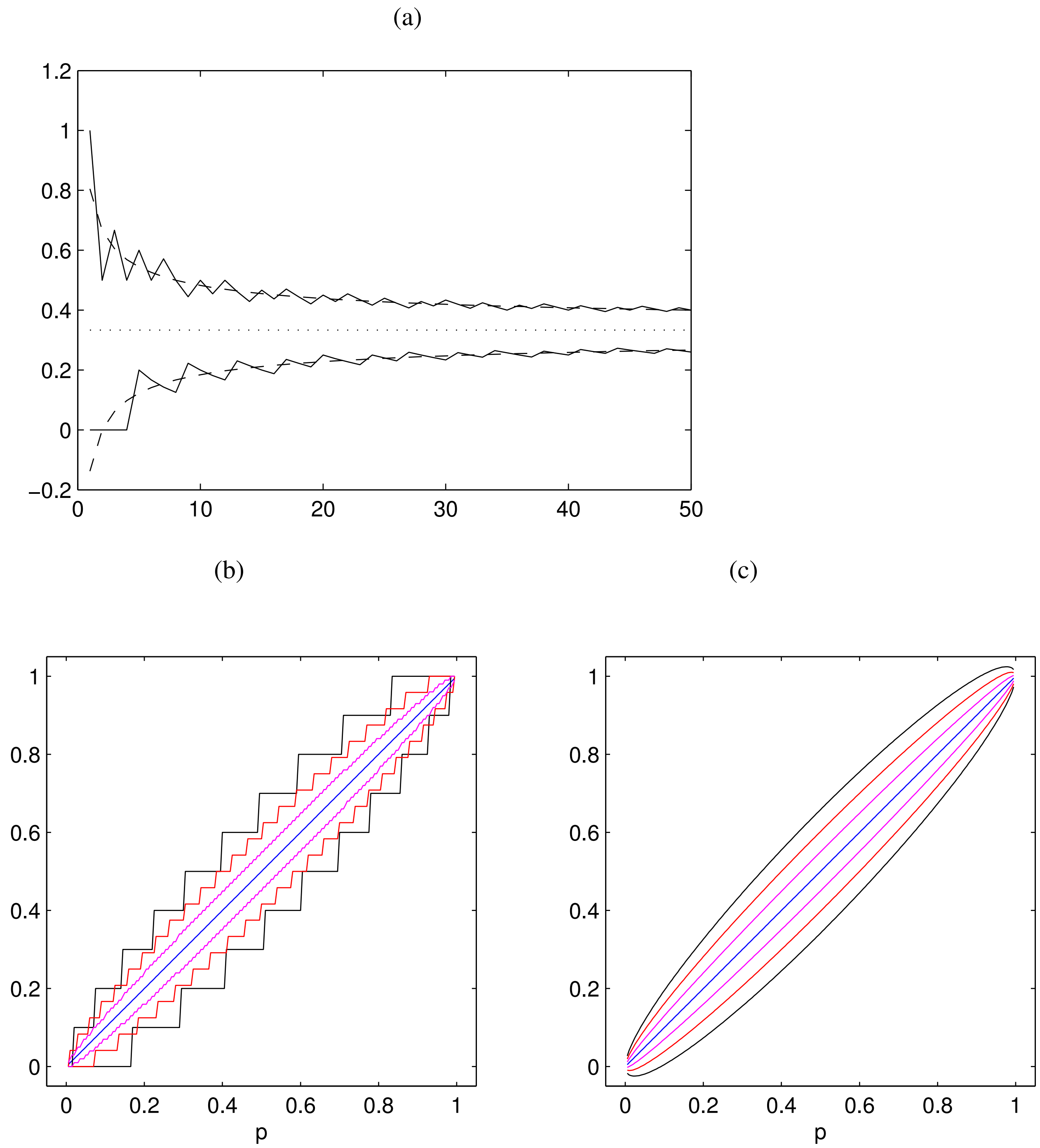


Figure 2. (a) 16th and 84th percentiles of the counting estimate p_N (solid lines) and p plus and minus the standard deviation of the estimate p_N (dashed line) for $p = 1/3$ (dotted line). Curves of (b) 16th and 84th percentile and (c) mean plus and minus the standard deviation as a function of the true tercile probability p for ensemble sizes $N = 10$ (black), 24 (red), 100 (magenta), and ∞ (blue).

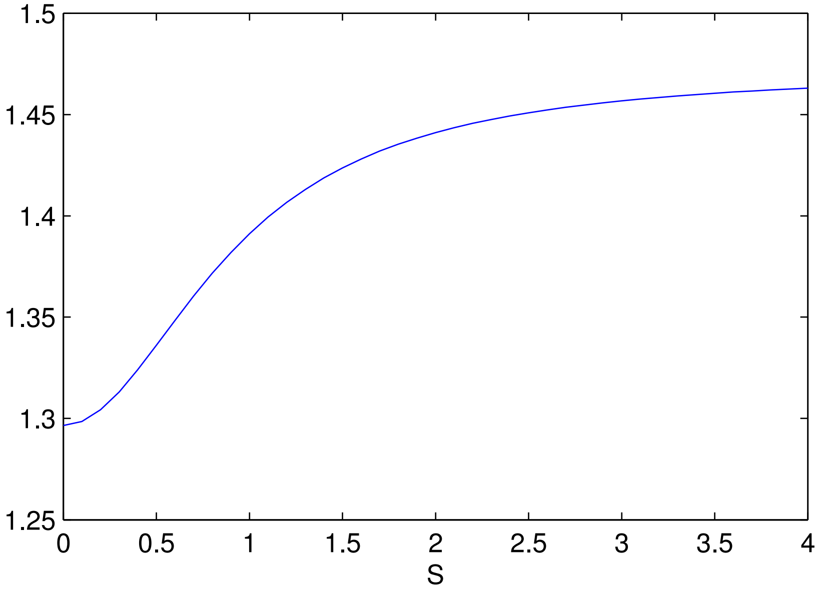


Figure 3. Ratio of the standard deviation of the counting and Gaussian fit estimates as a function of signal-to-noise ratio S.

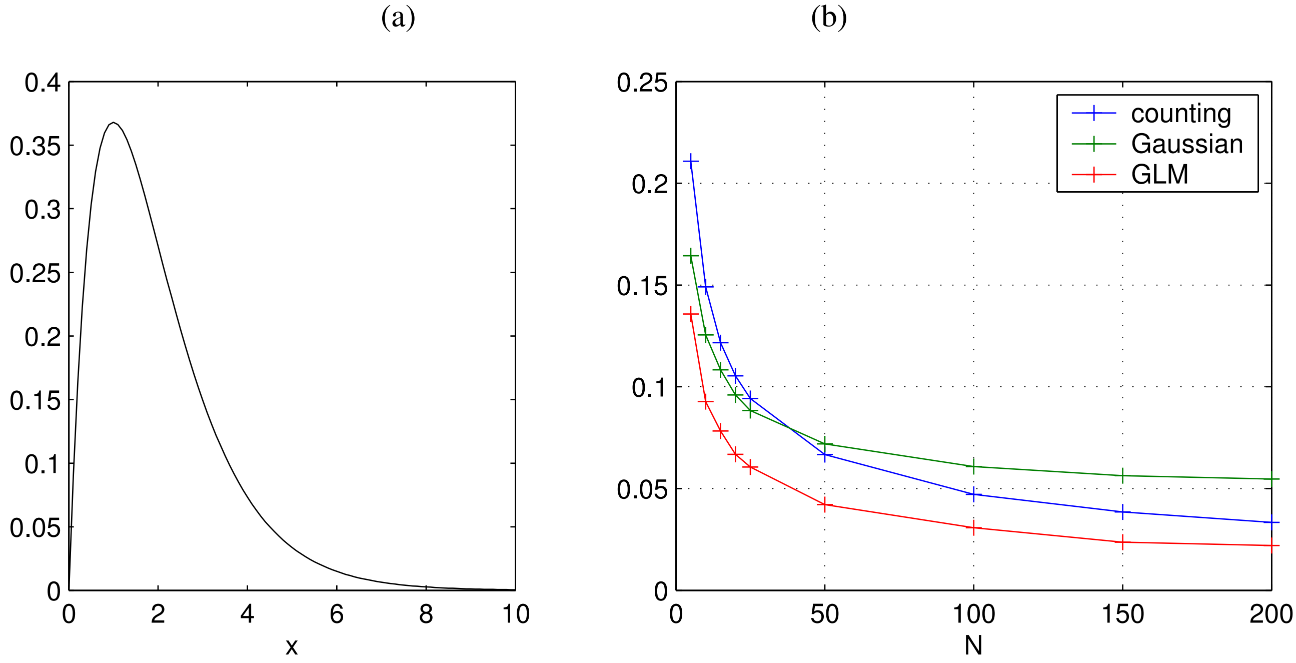


Figure 4. The (a) pdf and (b) rms error as a function of ensemble size N for the counting, Gaussian fit and GLM tercile probability estimates where the variable x has a gamma distribution with parameters (2,1).

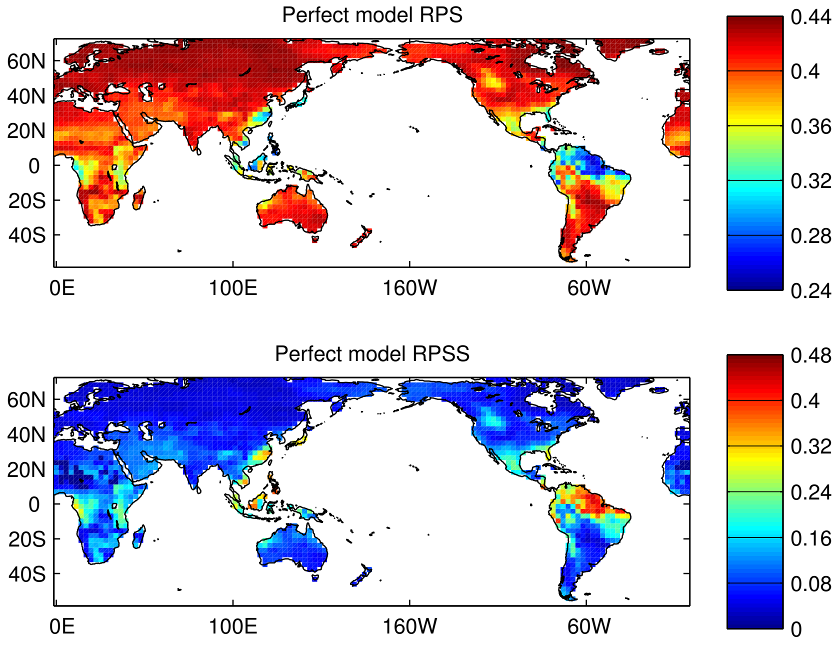


Figure 5. Perfect model (a) RPS and (b) RPSS.

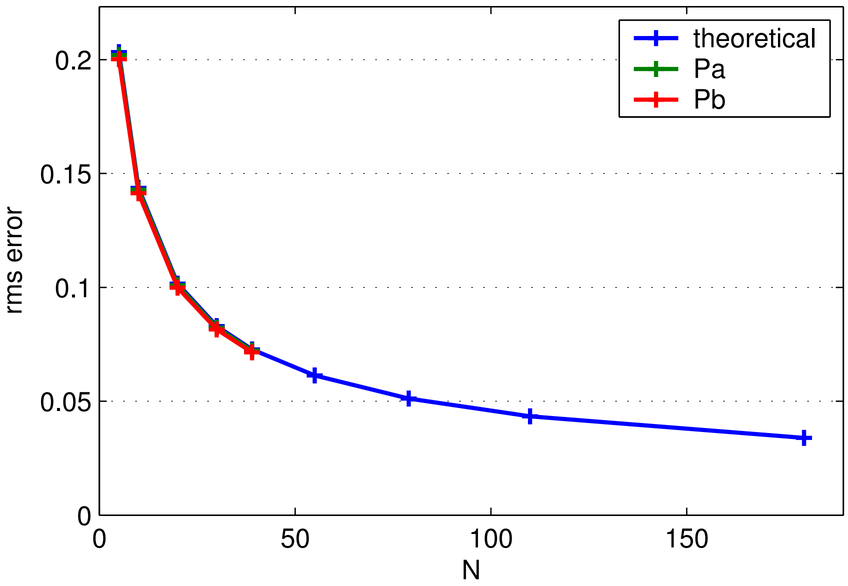


Figure 6. Gridpoint-average of the theoretical and empirically estimated standard deviation of the tercile probability estimate.

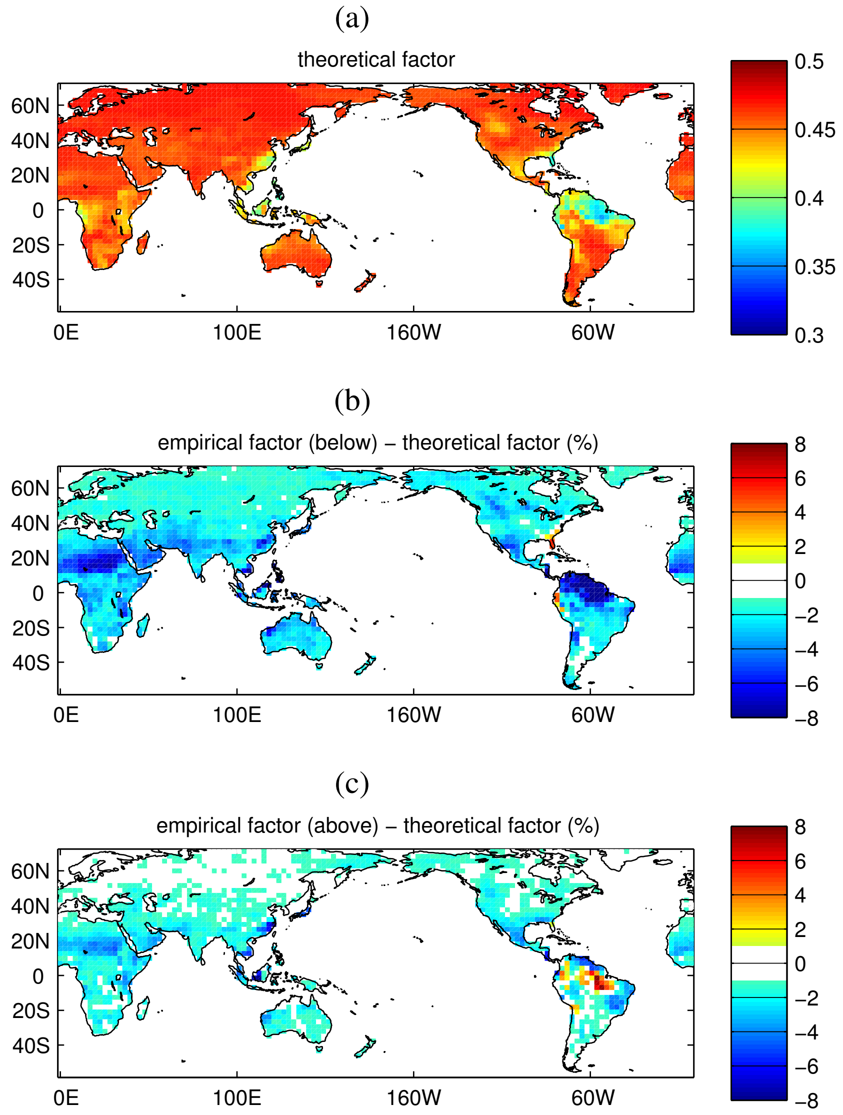


Figure 7. (a) Spatial variation of the quantity $\sqrt{2}/3/(1+s^2)^{1/4}$. Percent difference with the (b) below and (c) above sub-sampled estimates.

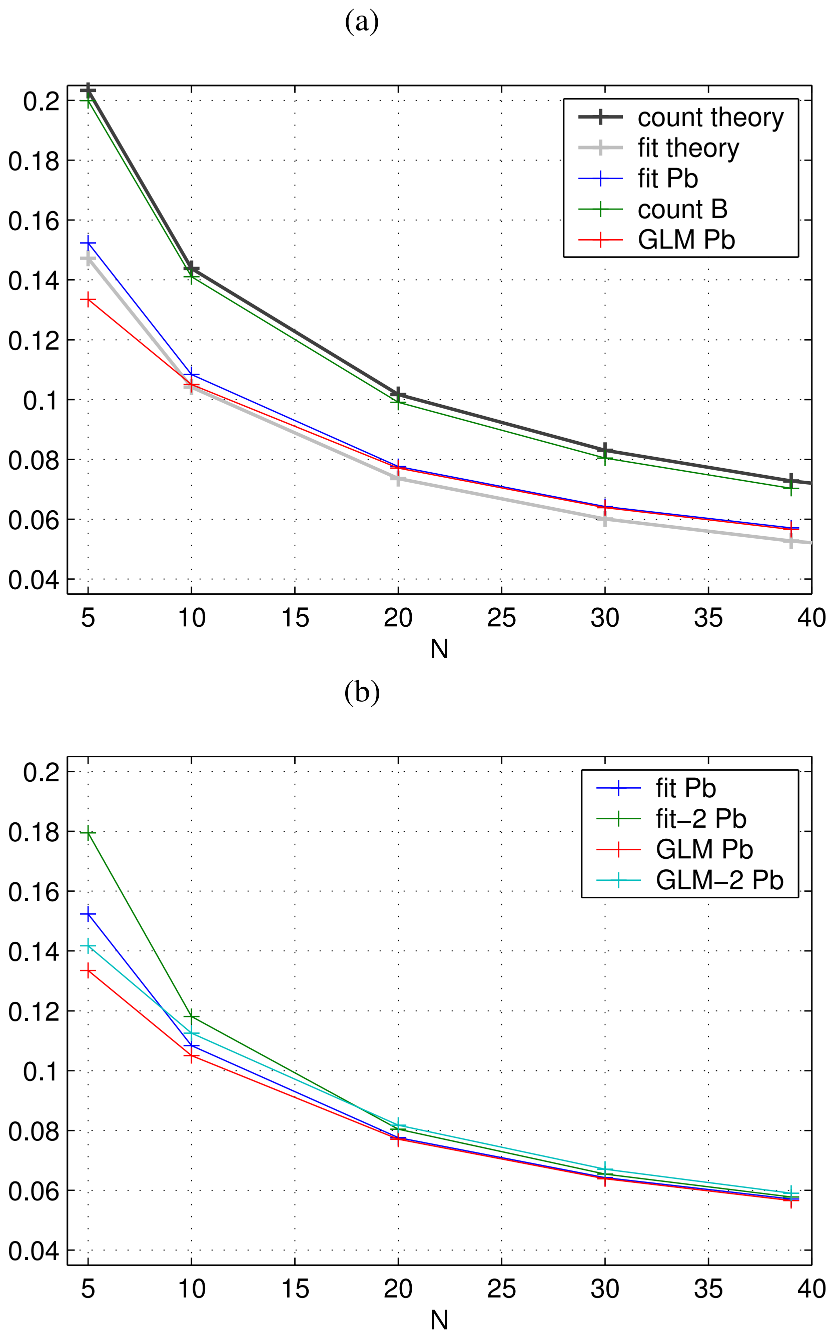


Figure 8. RMS error of the below-normal probability as a function of ensemble size N for the (a) 1-parameter and (b) 2-parameter estimates. The gray curves in panel (a) are the theoretical error levels for the counting and Gaussian fit methods. Fit-2 (GLM-2) denotes the two-parameter Gaussian (GLM) method.

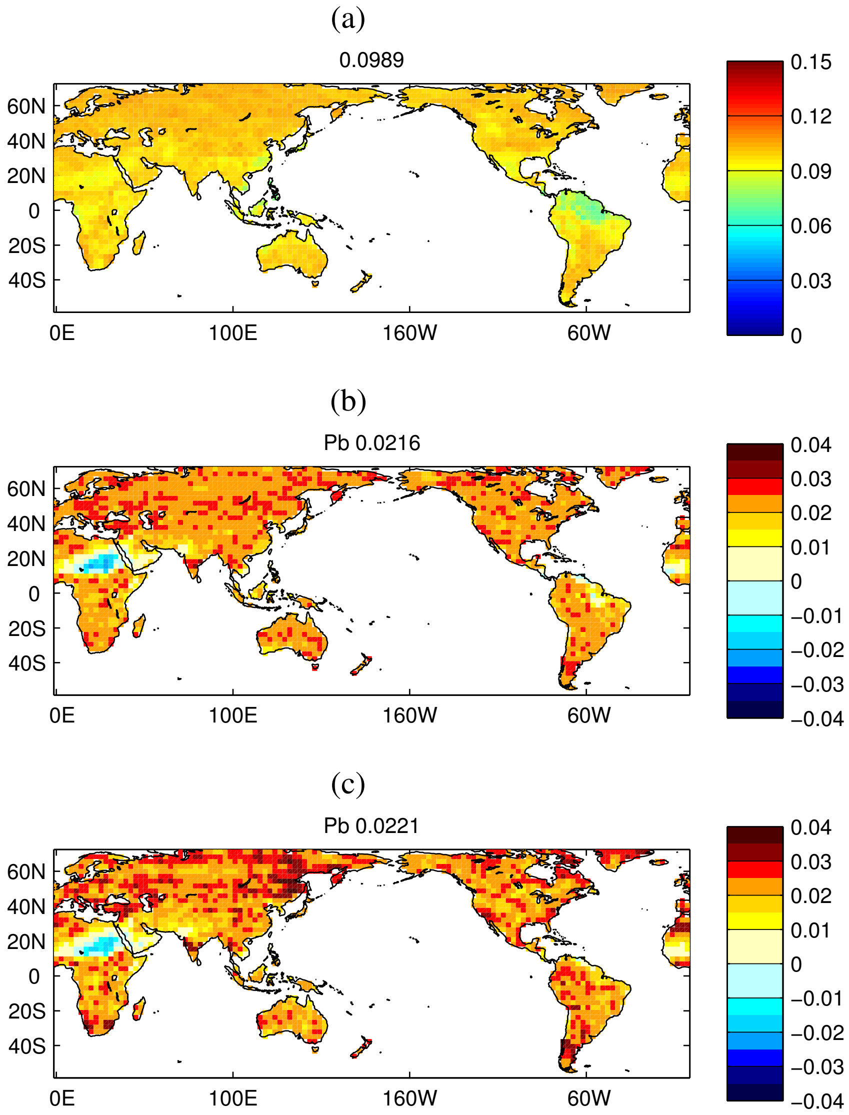


Figure 9. (a) RMS error of the counting estimate of the below-normal tercile probability with ensemble size 20. The RMS error of the counting error minus that of the (b) Gaussian fit and (c) the GLM based on the ensemble mean.

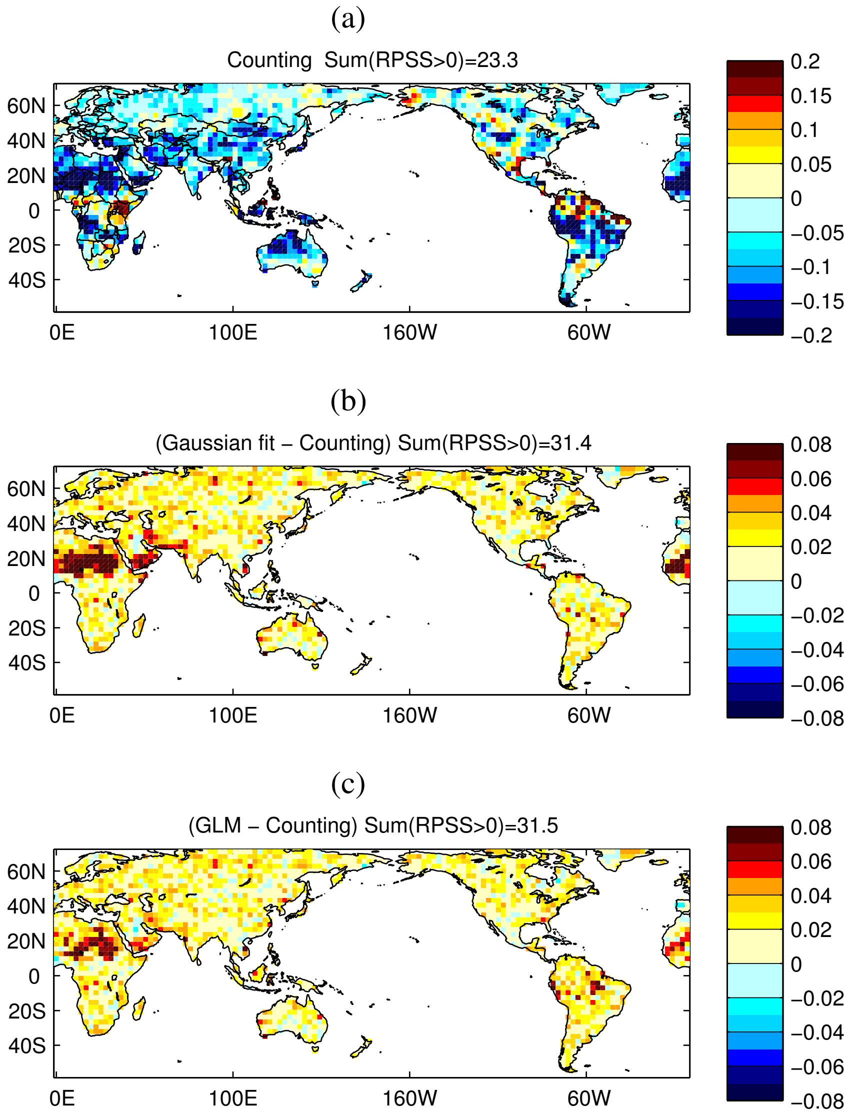


Figure 10. RPSS of (a) the counting-based probabilities and its difference with that of the (b) Gaussian and (c) GLM estimated probabilities. The sum of positive RPSS values is shown in the titles. Positive values in (b) and (c) correspond to increased RPSS compared to counting.

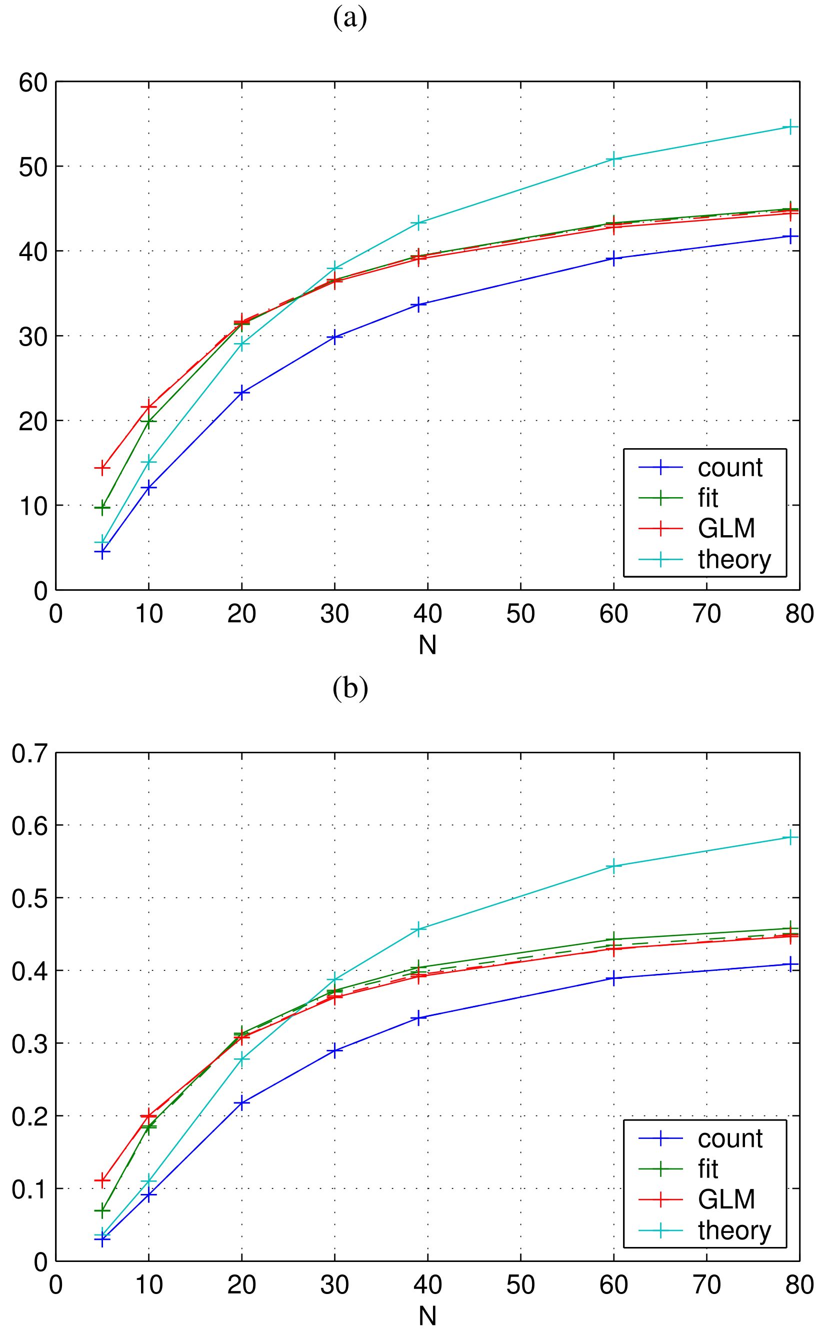
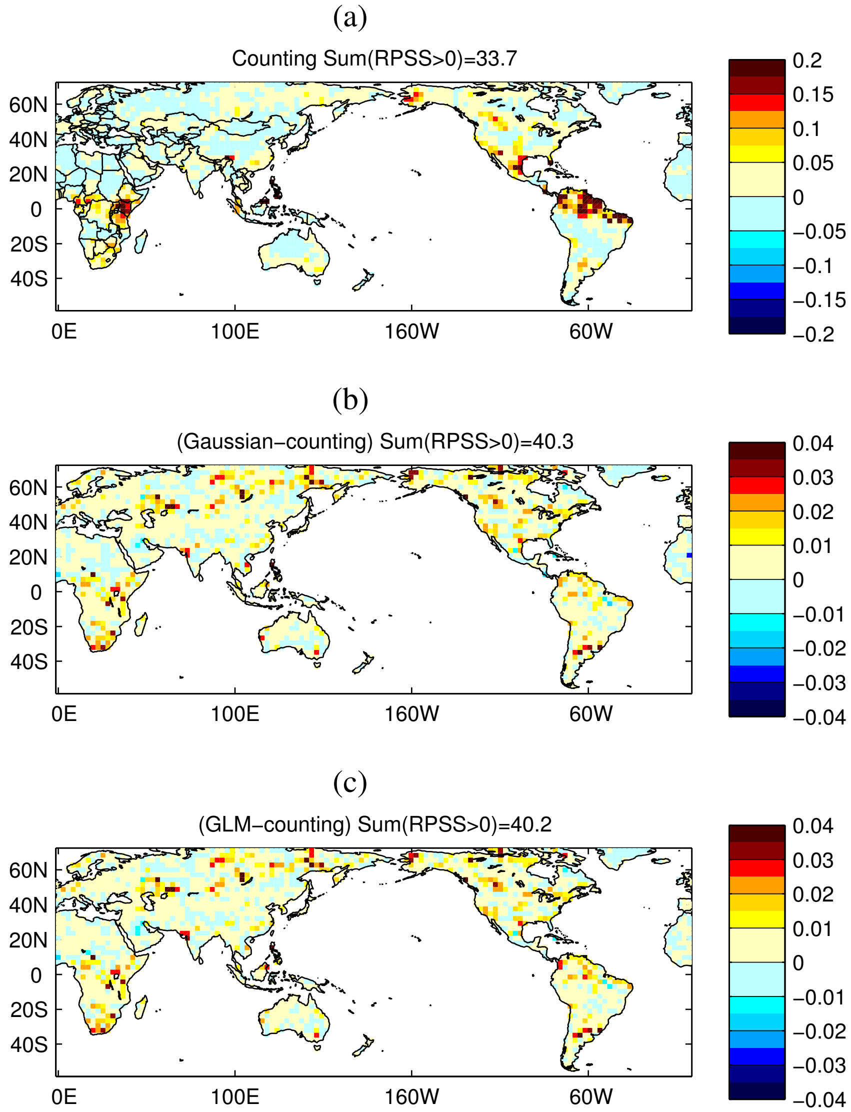


Figure 11. Gridpoint sum of (a) positive RPSS and (b) fraction of land points with RPSS> 0.



(a)

Counting Sum(RPSS>0)=33.7

(b)

(Gaussian−counting) Sum(RPSS>0)=40.3

(c)

(GLM−counting) Sum(RPSS>0)=40.2

Figure 12. As in Fig. 10 but for the Bayesian calibrated probabilities.

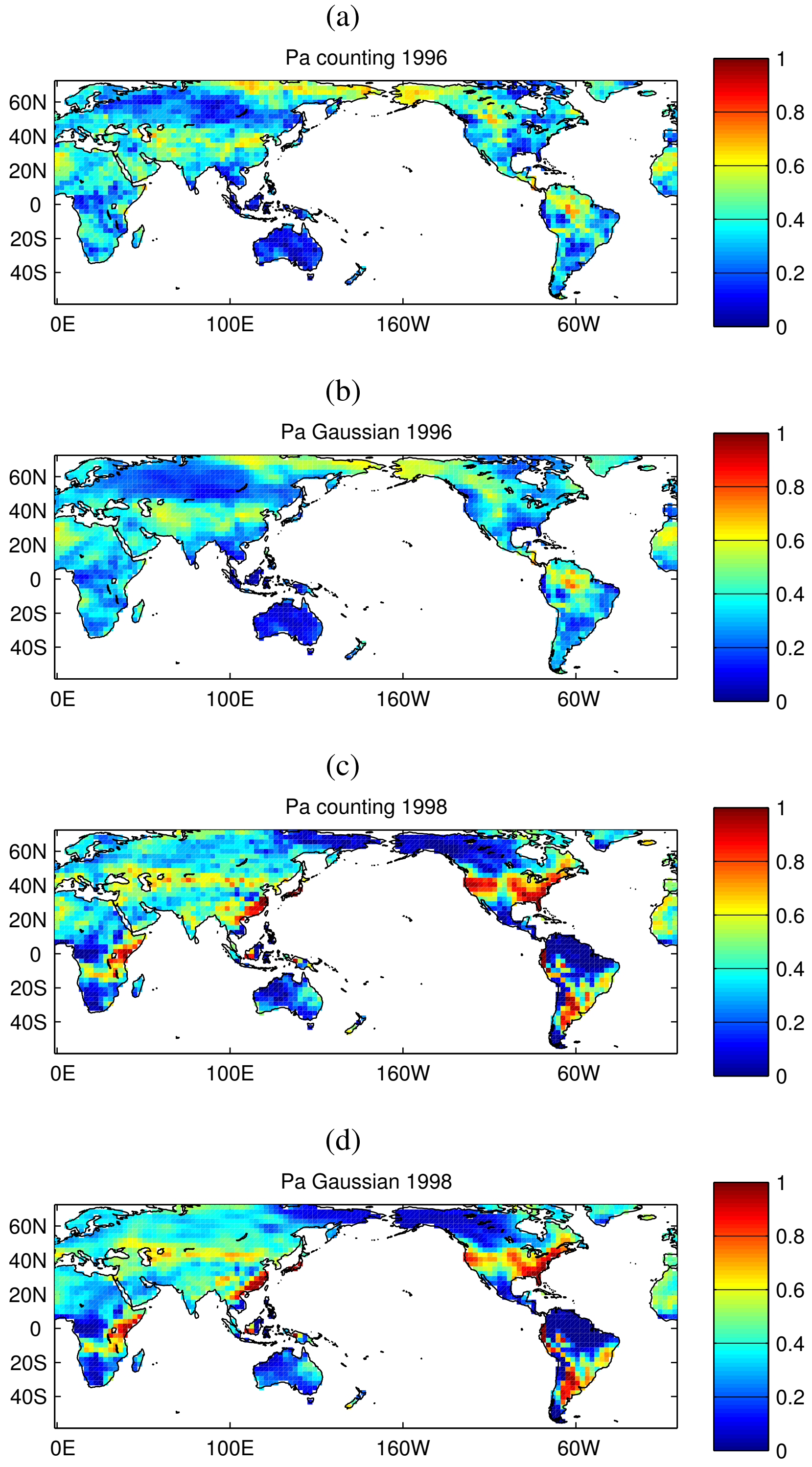


Figure 13. Probability of above-normal precipitation for DJF 1996 estimated by (a) counting and (b) Gaussian fit, and DJF 1998 using (c) counting and (d) Gaussian fit.