Rank Histograms of Stratified Monte Carlo Ensembles

STEFAN SIEGERT

Max Planck Institute for the Physics of Complex Systems, Dresden, Germany

JOCHEN BRÖCKER

Max Planck Institute for the Physics of Complex Systems, Dresden, Germany, and Visiting Research Fellow, Centre for the Analysis of Time Series, London School of Economics, London, United Kingdom

HOLGER KANTZ

Max Planck Institute for the Physics of Complex Systems, Dresden, Germany

(Manuscript received 14 October 2011, in final form 2 January 2012)

ABSTRACT

The application of forecast ensembles to probabilistic weather prediction has spurred considerable interest in their evaluation. Such ensembles are commonly interpreted as Monte Carlo ensembles meaning that the ensemble members are perceived as random draws from a distribution. Under this interpretation, a reasonable property to ask for is statistical consistency, which demands that the ensemble members and the verification behave like draws from the same distribution. A widely used technique to assess statistical consistency of a historical dataset is the rank histogram, which uses as a criterion the number of times that the verification falls between pairs of members of the ordered ensemble. Ensemble evaluation is rendered more specific by stratification, which means that ensembles that satisfy a certain condition (e.g., a certain meteorological regime) are evaluated separately. Fundamental relationships between Monte Carlo ensembles, their rank histograms, and random sampling from the probability simplex according to the Dirichlet distribution are pointed out. Furthermore, the possible benefits and complications of ensemble stratification are discussed. The main conclusion is that a stratified Monte Carlo ensemble might appear inconsistent with the verification even though the original (unstratified) ensemble is consistent. The apparent inconsistency is merely a result of stratification. Stratified rank histograms are thus not necessarily flat. This result is demonstrated by perfect ensemble simulations and supplemented by mathematical arguments. Possible methods to avoid or remove artifacts that stratification induces in the rank histogram are suggested.

1. Introduction

A forecast ensemble (or simply ensemble) is a collection of runs of a dynamical model. Heterogeneity of ensemble members is affected through different initial conditions, different model parameters, or the ensemble members can even be runs of different models for the same process. In any case, what makes the collection of runs an ensemble is their common target. That is, all members of the ensemble verify at the same time and are eventually compared to the same measurement: the verification. Forecast ensembles convey not only a best guess of the verification but also information about the uncertainty of that best guess. This uncertainty is not a physical property of the verification. It is rather the manifestation of incomplete knowledge of the forecaster about the initial state and the physics of the system. As a consequence, different forecasters might have different forecast uncertainties.

The forecaster's uncertainty about the verification is often conceptualized by a forecast distribution. The goal of an ensemble forecasting system is to use the sensitivity of the dynamical system to the imposed perturbations to generate a number of samples from the forecast distribution. However, the representation of the forecast distribution by the ensemble might be incorrect, that is, the verification might not behave like a draw from the forecast distribution that is estimated (or sampled) by the ensemble.

Corresponding author address: Stefan Siegert, Max-Planck-Institut für Physik komplexer Systeme, Nöthnitzer Str. 38, D-01187 Dresden, Germany.
E-mail: siegert@pks.mpg.de

DOI: 10.1175/MWR-D-11-00302.1

In this case, the uncertainty information in the ensemble is unwarranted, and thus its usefulness is limited.

If the forecast distribution is well reproduced by both the numerical model and the imposed perturbations, the ensemble members should be thought of as equally likely scenarios for the verification. An ensemble whose members are statistically indistinguishable from the verification is called statistically consistent (Anderson 1996). In a consistent ensemble, the rate at which the verification falls between any two adjacent ensemble members should be independent of the position of these ensemble members in the ordered ensemble. In other words, given that K ensemble members define $K + 1$ possible intervals into which the verification can fall, none of these intervals should be preferred by the verification and each interval should occur with an average relative frequency of $1/(K + 1)$. A histogram of the number of times that the verification falls into each interval in a historical dataset of forecast–verification pairs should then be flat, up to random fluctuations due to the finiteness of the number of samples. Such rank histograms (Talagrand et al. 1997; Hamill and Colucci 1997) are used to evaluate ensemble forecasts. Flatness of the rank histogram, or uniformity of the verification rank distribution, is considered a necessary condition for ensemble consistency.

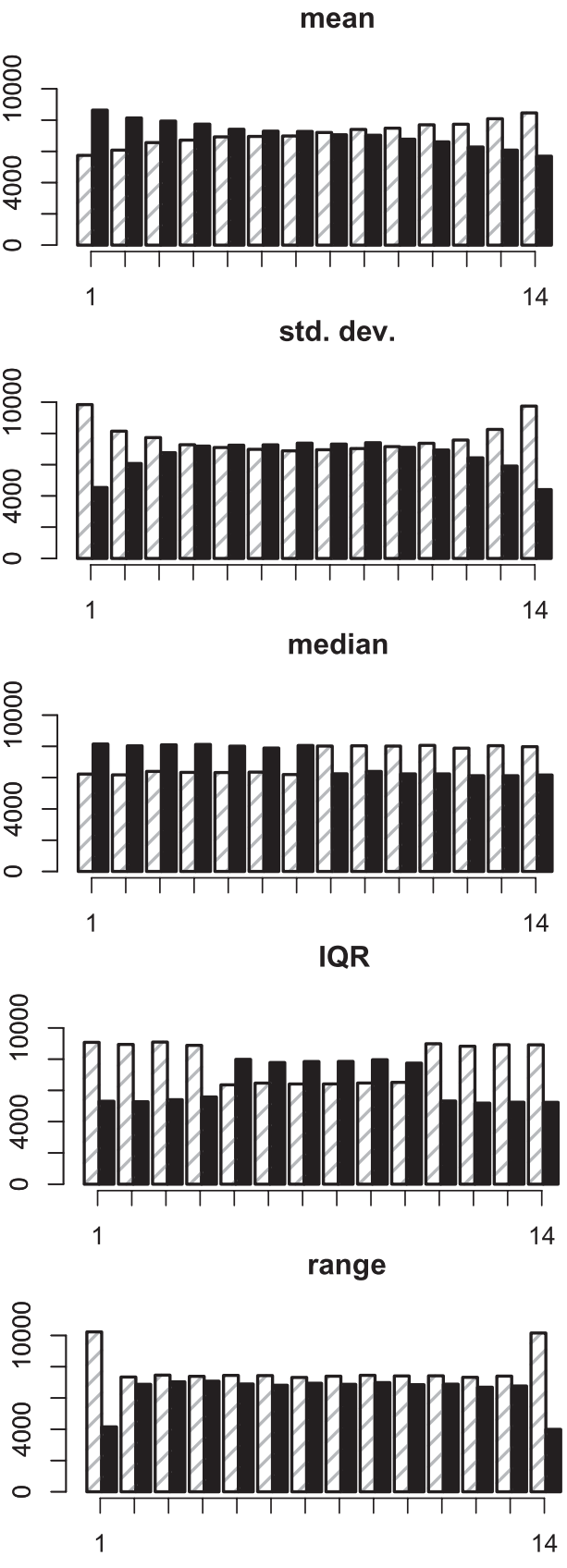
Some authors have made ensemble analyses more specific by means of ensemble stratification (Hamill and Colucci 1998; Bröcker 2008). The complete historical dataset of forecast–verification pairs was divided into subsets in which certain criteria are satisfied by the ensemble. Rank histograms constructed separately for each subset are used to assess whether the model reproduces the forecast distribution equally well (or badly) under the respective criteria. In Hamill and Colucci (1997) and Hamill and Colucci (1998), rank histograms of ensembles stratified along the ensemble standard deviation and along baroclinic instability (a function of the ensemble mean) are presented. Rank histograms for those ensembles with particularly low and high standard deviation or baroclinic instability are shown separately. Hamill (2001) advocates stratification along functions of the ensemble. Bröcker (2008) stratifies along the estimated ranked probability score of the ensemble, a function of the ensemble that correlates with the ensemble standard deviation. Siegert et al. (2011) stratify along the ensemble range and find that the ensemble range contains information about the occurrence of outliers. Further approaches to stratification in the literature include stratification along the verification (Mullen and Buizza 2002) and stratification along the current season (Peel and Wilson 2008; Siegert et al. 2011). All authors observe different rank histograms under different strata.

The application of rank histograms to stratified ensembles suggests that a flat rank histogram should be expected if the original (unstratified) ensemble is consistent, even though none of the authors cited above state this explicitly. In general, however, this assumption does not hold. In Fig. 1, stratified rank histograms of a dataset of perfectly consistent ensemble–verification pairs are shown. This plot readily demonstrates that stratification might turn a consistent ensemble into several inconsistent ones. To understand why that happens, consider as an example stratification along the mean of a hypothetical temperature ensemble, as in the top panel of Fig. 1. There are two effects that lead to variations in the ensemble mean: on the one hand, the physics of the system lead to warm and cold regimes under which ensembles have particularly high and low mean values, respectively. On the other hand, random sampling fluctuations cause the sample mean to be either warmer or colder than the true mean. This latter effect is purely random over the entirety of ensembles drawn from their distributions. However, by stratification, ensembles with an anomalously low sample mean are separated from ensembles with an anomalously high sample mean. The random sampling error is turned into a systematic error by stratification. In the ensembles with an anomalously low sample mean, the verification drawn from the same distribution has an increased tendency to fall into the higher ranks, and vice versa. The rank histograms of these subsets of the original ensemble are then sloped.

The new contributions of the present study are a detailed description of the effects of stratification on the rank histogram, and the introduction of possible methods to cope with these effects. To this end, ensemble forecasting is formalized in section 2. A connection between forecast ensembles, the Dirichlet distribution, and sampling from the probability simplex is established. In section 3, the theory of ensemble stratification is reviewed. In section 4, a perfect ensemble study is presented that highlights the effects of stratification under different criteria. In section 5, mathematical arguments based on the preceding sections are provided in order to explain the effects of stratification. Section 6 presents a statistical test for ensemble consistency under stratification along certain criteria, and section 7 proposes methods to avoid artifacts induced by stratification. Section 8 concludes with a discussion and summary. Readers only interested in a phenomenological description of the problem may proceed to section 4 right away.

2. Monte Carlo ensembles and the probability simplex

A forecast ensemble drawn randomly and independently from a distribution can be thought of as arising in



the following way. Denote the forecast distribution at instance t as $F_t(\cdot)$, here taken as a cumulative distribution function (cdf). Then $F_t(x)$ is equal to the forecast probability concentrated in the interval $(-\infty, x]$ at time instance t. It should be kept in mind that F_t generally changes over time in forecasting problems because the degree of uncertainty of the forecaster about the future is different on different occasions. Consider further a K-sample $\mathbf{u} = (u_{[1]}, \ldots, u_{[K]})$ drawn uniformly and independently from the unit interval and ordered by increasing magnitude—$u_{[i]}$ denotes the ith-order statistic. Then an ordered K-member ensemble $\mathbf{e}$ drawn independently from F_t can be constructed by evaluating the inverse of F_t at $\mathbf{u}$:

$$\mathbf{e} = \{F_t^{-1}(u_{[1]}), \ldots, F_t^{-1}(u_{[K]})\} \quad (1)$$

$$= (e_{[1]}, \ldots, e_{[K]}). \quad (2)$$

This transformation is a result of the fact that if x is drawn from the cdf $F(\cdot)$, then the probability integral transform (PIT) of x, $F(x)$, is uniformly distributed on the unit interval (Mood et al. 1974). We call ensembles that are constructed in this way Monte Carlo ensembles (MCEs). The ordering between the $u_{[i]}$ and $e_{[i]}$ is preserved by the PIT because $F_t(\cdot)$ is a monotonically increasing function. In operational ensemble forecasting it is usually assumed that ensemble members behave like independent samples drawn from a forecast distribution. That is, operational ensembles are usually interpreted as MCEs.

An MCE is statistically consistent if the verification is statistically indistinguishable from the ensemble members (i.e., if it can be considered a random independent draw from the forecast distribution). In other words, under statistical consistency every ensemble member can be considered an equally likely scenario for the verification under the uncertainty of the forecaster.

Given an MCE, it is a relevant question to ask what the probability is that the verification η will fall between an adjacent pair of ensemble members $e_{[i-1]}$ and $e_{[i]}$ for $i = 1, \ldots, K + 1$, where we define $e_{[0]} = -\infty$ and $e_{[K+1]} = +\infty$. If we assume that the ensemble is statistically consistent, then this probability is given by

←

FIG. 1. Patterns in the rank histogram that result from stratification along different sample statistics. Abscissas indicate verification ranks and ordinates indicate frequency of occurrence. Dark (light hatched) bars correspond to the high (low) stratum. None of these rank histograms can be considered flat, which has been confirmed by a χ^2 test.

$$q_i := \mathbb{P}\{\eta \in (e_{[i-1]}, e_{[i]}] \mid \mathbf{e}, F_t\} \tag{3}$$

$$= F_t(e_{[i]}) - F_t(e_{[i-1]}) \tag{4}$$

$$= F_t\{F_t^{-1}(u_{[i]})\} - F_t\{F_t^{-1}(u_{[i-1]})\} \tag{5}$$

$$= u_{[i]} - u_{[i-1]}. \tag{6}$$

Hence, given the forecast distribution F_t, each MCE defines a $(K + 1)$-dimensional random vector given by

$$\mathbf{q} := (q_1, \ldots, q_{K+1}) \tag{7}$$

$$= (u_{[1]}, u_{[2]} - u_{[1]}, \ldots, u_{[K]} - u_{[K-1]}, 1 - u_{[K]}). \tag{8}$$

The probability of the verification falling between ensemble members $e_{[i-1]}$ and $e_{[i]}$ is equal to q_i. If the verification does indeed fall into this interval, we say that it has *rank i*. For this reason, q_i can be called the rank probability.

The set of J-dimensional vectors whose elements are nonnegative and sum to one is called the $(J - 1)$-dimensional *probability simplex*. It is a $(J - 1)$-dimensional surface embedded in J dimensions. Every vector $\mathbf{q} = (q_1, \ldots, q_J)$ that lies on the $(J - 1)$-dimensional probability simplex can be interpreted as a J-dimensional probability mass function (pmf), that is, a mutually exclusive and collectively exhaustive probability assignment over J categories. (The two-dimensional probability simplex is visualized in Figs. 2 and A1.)

The joint density of the elements of the K-dimensional vector $\mathbf{u} = (u_{[1]}, \ldots, u_{[K]})$ is given by

$$p_{u_{[1]},\ldots,u_{[K]}}(a_1, \ldots, a_K) = K! = \text{const}, \tag{9}$$

if $0 < a_1 < \cdots < a_K < 1$ and zero otherwise (Mood et al. 1974). The transformation given by Eq. (8), which maps the vector $\mathbf{u}$ to the vector $\mathbf{q}$ (which is a pmf), is a linear transformation. It follows that the joint density of the elements of $\mathbf{q}$ is the same as that of the elements of $\mathbf{u}$, up to a constant (Mood et al. 1974). Thus, the joint density of the elements of $\mathbf{q}$ is likewise uniform if $\mathbf{q}$ is an element of the probability simplex, and zero otherwise. We conclude that the $(K + 1)$-dimensional rank probability vector $\mathbf{q}$ given by Eq. (8) can be considered a random pmf, drawn uniformly from the K-dimensional probability simplex.

The uniform distribution on the probability simplex is part of a larger family known as the Dirichlet distributions (Frigyik et al. 2010) which we review in appendix A. Uniform sampling from the simplex implies that the $K + 1$ parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{K+1})$ of the Dirichlet distribution $\text{Dir}(\boldsymbol{\alpha})$ are constant and equal to one. That is, for a random pmf $\mathbf{q}$ whose elements are defined by a statistically consistent MCE as in Eq. (3), we have $\mathbf{q} \sim \text{Dir}(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} = (1, \ldots, 1)$.

An important consequence of this Monte Carlo interpretation of forecast ensembles is that the rank probability is not a constant equal to $1/(K + 1)$ even if the ensemble is statistically consistent. Rather, it is subject to fluctuations due to the random sampling of the ensemble members. It is a random variable distributed like the marginal of a Dirichlet distribution (i.e., a beta distribution with parameters 1 and K; Frigyik et al. 2010). Depending on the ensemble, certain verification ranks become more or less likely at a given instance. It is merely the expectation value of the q_i that satisfies

$$\mathbb{E}[q_i] = \frac{1}{K + 1}, \tag{10}$$

according to Eq. (A4).

An established method to evaluate forecast ensembles is the verification rank histogram (Talagrand et al. 1997; Hamill and Colucci 1997), which is a histogram of verification ranks in a historical dataset of ensemble forecasts. The expected height of the ith histogram bar is governed by the expectation of the corresponding rank probability $\mathbb{E}[q_i]$, which is given by $1/(K + 1)$ for all i in a consistent MCE. A flat histogram is usually taken as a necessary condition for a consistent forecast ensemble. In practice, rank histograms are often U shaped or sloped, indicating lack of variability or unconditional bias of the forecast ensemble, respectively (Hamill 2001). In view of the considerations in the present section, a single verification rank can be interpreted as a single draw from a random $(K + 1)$-dimensional pmf which, in turn, was drawn from a Dirichlet distribution with parameters $\boldsymbol{\alpha} = (1, \ldots, 1)$. The rank histogram is thus a summary of N individual draws from random pmfs.

3. Ensemble stratification

In the following we focus on a practical domain, namely, ensemble stratification (Bröcker 2008), in which the results presented so far are relevant. Ensemble stratification amounts to imposing a certain stratification criterion under which a historical dataset of ensemble forecasts is partitioned. Stratification partitions the original ensemble into two or more strata.

Ensemble stratification has the benefit of facilitating more refined and flow-dependent consistency analyses of a forecast ensemble. For example, a forecaster might want to assess the consistency of his temperature ensemble only under warm conditions. For this purpose he can partition the ensembles into a warm and a cold stratum

and analyze the warm stratum individually. Finding different systematic inconsistencies under different strata is valuable from a diagnostic point of view because then more specific corrections can be applied to the forecast.

The question is then how to identify "warm conditions," or more generally, how to realize stratification. One possible group of stratification criteria are parameters of the underlying forecast distribution, such as its mean or variance. Such criteria are considered in Bröcker (2008). If the forecast model indeed produces ensemble members that are indistinguishable from the verification, rank histograms of the individual strata should still be flat. This can be seen by writing the forecast distribution as $F_\kappa(\cdot)$ and assuming for a moment that the verification is drawn from a distribution $F_\theta(\cdot)$. Assume that the two distributions are from the same family of distributions and that the difference between them as well as their time dependencies, if any, are completely contained in the parameters θ and κ. The distributions are supposed to be identical if and only if $\theta = \kappa$. If the verification η is a draw from F_θ and the members $e_{[i]}$ of the ensemble $\mathbf{e}$ are drawn from F_κ, the probability q_i that the verification falls between an adjacent pair of ensemble members $e_{[i-1]}$ and $e_{[i]}$ is given by

$$q_i = \mathbb{P}\{\eta \in (e_{[i-1]}, e_{[i]}] \,|\, \mathbf{e}, F_\theta, F_\kappa\} \tag{11}$$

$$= F_\theta(e_{[i]}) - F_\theta(e_{[i-1]}) \tag{12}$$

$$= F_\theta\{F_\kappa^{-1}(u_{[i]})\} - F_\theta\{F_\kappa^{-1}(u_{[i-1]})\}. \tag{13}$$

Assume $\kappa \in \mathbb{K}$. Stratification amounts to defining a function $S(\cdot)\colon \mathbb{K} \to (1, \ldots, M)$ that maps the parameter $\kappa \in \mathbb{K}$ of the forecast distribution to one of M discrete indices. The expected height of the bars in a rank histogram is determined by the expectation of the rank probability q_i. Under the mth stratum, that is when $S(\kappa) = m$, this expectation is

$$\mathbb{E}[q_i \,|\, S(\kappa) = m] \tag{14}$$

$$= \mathbb{E}[F_\theta\{F_\kappa^{-1}(u_{[i]})\} - F_\theta\{F_\kappa^{-1}(u_{[i-1]})\} \,|\, S(\kappa) = m]. \tag{15}$$

If the ensemble is statistically consistent under the mth stratum, then $\theta = \kappa \;\forall\; \{\kappa\colon S(\kappa) = m\}$, and Eqs. (14)–(15) are equal to

$$\mathbb{E}[u_{[i]} - u_{[i-1]} \,|\, S(\kappa) = m] = \frac{1}{K+1}. \tag{16}$$

Equation (16) shows that if the ensemble is statistically consistent under the mth stratum, the expectation of q_i conditional on the stratification criterion is equal to the expectation of $u_{[i]} - u_{[i-1]}$ conditional on the criterion. Since the $u_{[\cdot]}$ are uniformly distributed, they are independent of κ and thus the expectation in Eq. (16) is independent of conditioning on κ. If, on the other hand, $S(\theta) \neq S(\kappa)$ under some stratum m, Eq. (15) is not equal to Eq. (16) for that stratum. Therefore $\mathbb{E}[q_i \,|\, S(\kappa) = m]$ is indeed conditional on the stratification criterion and thus not necessarily equal to $1/(K+1)$. The rank histogram under the mth stratum is then not flat.

If the number of instances in the historical dataset in which the ensemble falls into the mth stratum is small compared to the entirety of all instances, a systematic deviance of $\mathbb{E}[q_i \,|\, S(\kappa) = m]$ might remain undetected by a rank histogram that is constructed over all instances. Considering only the respective subset of ensembles by means of stratification increases the detectability of such a systematic misrepresentation of the forecast distribution.

In practice though, there is a difficulty with this approach to stratification. The forecast distribution, along whose parameters the ensemble is stratified, is usually not available to the forecaster. In numerical weather forecasting, the forecaster has the possibility to sample from that distribution by running a number of simulations; the distribution itself though is not available to him in closed form. Hence, stratifying along a parameter of the distribution might be infeasible in practice. But since parameters of a distribution can be estimated from a random sample drawn from that distribution, the forecaster might instead stratify along estimates of these parameters calculated from the ensemble. Alternatively, one could stratify ensembles along functions of the ensemble members in general. However, we will show in the following two sections that such an approach leads to inconsistent strata.

4. Nonflat rank histograms due to stratification

In this section we consider stratification criteria that are functions of the ensemble that is analyzed. Examples for such criteria are parameter estimates of the underlying distribution, such as the ensemble mean or variance. The rank histograms of these strata are in general not flat, even if the original unstratified ensemble is perfectly consistent. To demonstrate this we present results of a perfect ensemble simulation. More mathematical arguments follow in section 5.

We produce a dataset of a consistent forecast ensemble by the following procedure. 1) Two numbers μ and σ are sampled randomly and uniformly from the intervals $[-1, 1]$ and $[1, 2]$, respectively, and taken as the mean and standard deviation of a Gaussian distribution $p(x \mid \mu, \sigma)$. 2) Then $K = 13$ random samples are drawn from this Gaussian $p(x \mid \mu, \sigma)$ and taken as ensemble members. 3) Another independent draw from $p(x \mid \mu, \sigma)$ is taken as the verification.

The procedure in steps 1–3 is repeated N times, each time with a different realization of μ and σ, which provides a dataset of N ensemble–verification pairs. An ensemble sampled in this manner is statistically consistent since ensemble members and verification are drawn from the same distribution $p(x \mid \mu, \sigma)$ at each instance. Ensemble data with such statistical behavior can arise in a temperature anomaly ensemble. The varying mean value of the distribution models the predicted intensity of the anomaly, and the varying standard deviation can be regarded as varying levels of forecast uncertainty.

After generating the dataset of ensemble–verification pairs, the ensemble is stratified in the following way. Here we take the ensemble standard deviation as the stratification criterion. The ensemble standard deviation is calculated for all N ensembles in the dataset. The empirical average over all N ensemble standard deviations in the dataset is calculated and taken as the threshold for stratification into two strata. All those ensemble–verification pairs in the dataset with an ensemble standard deviation falling below the threshold are collected in the "low" stratum, and those instances where the ensemble has an above-average ensemble standard deviation are put into the "high" stratum. Finally, a rank histogram is constructed for both strata separately.

Figure 1 shows rank histograms of such consistent forecast ensembles stratified along various criteria. These criteria are statistics calculated from the ensemble, including the ensemble mean, ensemble standard deviation, ensemble median ($Q_{0.5}$, the 0.5 quantile), ensemble interquartile range (IQR; $Q_{0.75}$–$Q_{0.25}$), and the total range between largest and smallest ensemble member. We have used a large value of $N = 2 \times 10^5$ in order to emphasize the systematic effects of stratification.

Obviously, none of the rank histograms in Fig. 1 can be considered flat. This has been confirmed by a χ^2 test (Pearson 1900, see also section 5). The effect of stratification along statistics of the ensemble members is rather to produce inconsistent strata. For example, stratification along the ensemble mean leads to sloped histograms while stratification along the median leads to a pronounced step between the central ranks. Note that the sum of the dark and light bars in each plot of Fig. 1 always yields the same rank histogram, namely, that of the original unstratified ensemble. This rank histogram can be considered flat, which we have checked with a χ^2 test.

An intuitive explanation as to why stratified rank histograms as the ones in Fig. 1 are not flat is as follows. Consider the ensemble mean, for example. Even if the ensemble is drawn from the same distribution at each instance, the ensemble mean will not exactly equal the true mean of the distribution; as a result of sampling fluctuations, some ensembles will have a smaller mean, and some will have a greater mean than the true distribution mean. Stratification separates these two groups from each other. In the "low mean" ensembles though, the verification, if drawn from the same distribution, is more likely to occupy a higher verification rank than in the "high mean" ensembles, leading to a sloped rank histogram. In this case, stratification separates a consistent ensemble into two strata of ensembles that underforecast and overforecast, respectively. Similar qualitative explanations can be given for the other stratification criteria presented in Fig. 1.

We have conducted a large number of perfect ensemble simulations using different stratification criteria and different distributions. A point worth mentioning is that the outer plateaus under stratification along range and IQR need not be of equal height as is the case in Fig. 1. If the distribution from which ensemble and verification are sampled is skewed, the heights of the outer plateaus are different from each other. Furthermore, a note on calculation of median and IQR might be in order. In statistical packages, different methods exist for approximating distribution quantiles from finite samples (Hyndman and Fan 1996). Some of these methods use interpolation techniques that take a weighted average of two samples to approximate a single quantile. Throughout this study we use a single ensemble member to approximate a single quantile. While introducing a certain bias into the quantile estimate, stratification along a single ensemble member instead of a superposition of several members significantly facilitates calculations as will become evident in section 5. Last, we should recall once again that such patterns do not emerge if the true distribution parameters are used instead of their estimates obtained from the consistent ensemble. This is so because the expectations of the rank probabilities are then independent of the stratification criterion, as was shown in section 3.

5. A formal description of the effect of ensemble stratification

To control and possibly correct for the behavior observed in section 4, we describe the phenomenon in more mathematical terms. We especially consider the case of stratifying a consistent ensemble along a single ensemble member as well as along the difference between two ensemble members. We provide a complete description of the rank histogram of an ensemble stratified along its kth largest ensemble member.

Let $\mathbf{e} = (e_{[1]}, \ldots, e_{[K]})$ denote a consistent forecast ensemble, drawn from a cdf $F(\cdot)$, with members in ascending order. Furthermore, let the stratification criterion $S(\mathbf{e})$: $\mathbb{R}^K \rightarrow \{1, 2\}$ be a function that maps the

ensemble **e** into one of two groups. Note that the discussion is simplified by assuming that S is only a function of the kth largest ensemble member $e_{[k]}$ [i.e., $S(\mathbf{e}) = S(e_{[k]})$]. We define a threshold $\tilde{\tau}$ such that $S(e_{[k]}) = 1$ if $e_{[k]} < \tilde{\tau}$ (low stratum) and $S(e_{[k]}) = 2$ otherwise (high stratum). Imposing an upper bound on $e_{[k]}$ is equivalent to imposing an upper bound on its PIT (see section 2) $F(e_{[k]})$ (i.e., on the mass of probability concentrated to the left of $e_{[k]}$). The PIT of $e_{[k]}$ is given by $\sum_{j=1}^{k} q_j$, where q_i is the ith rank probability as defined in Eq. (3). Under this setting, the height of the ith bar in the ensemble's rank histogram is governed by the expectation of the ith element of the rank probability vector **q**, conditional on the ensemble being in the mth stratum. Thus, in the low stratum [where $S(e_{[k]}) = 1$] we have

$$\mathbb{E}[q_i \,|\, e_{[k]} < \tilde{\tau}] = \mathbb{E}\left[q_i \Big| \sum_{j=1}^{k} q_j < \tau\right], \tag{17}$$

where $\tau = F(\tilde{\tau})$.

An upper bound τ on the sum over q_j is clearly also an upper bound τ on each of the q_j individually. Because of this upper bound, those q_j with $j \leq k$ are on average smaller than they would be without stratification. As a result, the q_j with $j > k$ must be larger on average such that the elements of $\mathbb{E}[\mathbf{q}]$ still sum to one.

This situation is illustrated in Fig. 2 for the case of $\dim(\mathbf{q}) = 3$ (i.e., $K = 2$) and an upper bound on the largest ensemble member $e_{[2]}$. Placing an upper bound on $e_{[2]}$ due to stratification is equivalent to placing an upper bound on the PIT of $e_{[2]}$, which is equal to $q_1 + q_2$. The upper bound on q_1 and q_2 limits the area on the two-dimensional probability simplex from which samples can be drawn. As a result, the expectation values of q_1 and q_2 become smaller than they would be without stratification while the conditional expectation of q_3 becomes larger.

Next, consider stratification by placing an upper bound on the difference $e_{[k]} - e_{[j]}$ between two ensemble members, where $k > j$. This is the case if we stratify along the IQR or ensemble range, for example. Constraining the difference between two ensemble members amounts to constraining the mass of probability concentrated between these two members. This upper bound on the enclosed mass of probability $\sum_{i=j+1}^{k} q_i$ translates into an upper bound on all the individual q_i with $i \in [j+1, k]$. As a result, the bars in the rank histogram corresponding to these indices will be lower than $1/(K+1)$. Accordingly there are now two steps in the rank histogram, one at each index corresponding to the ensemble members that enter the stratification criterion. By applying the same reasoning to three or more ensemble members we speculate that there should be a step in the rank histogram at every index whose corresponding member enters the criterion.

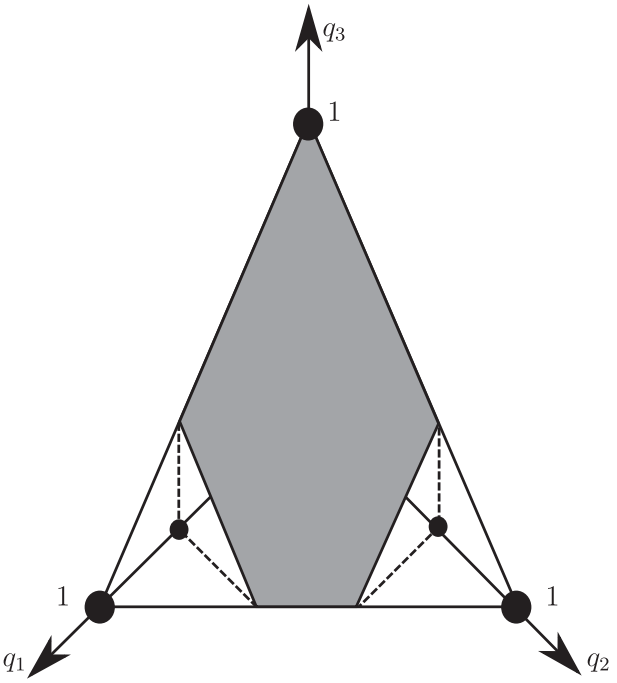


FIG. 2. The area defined by connecting the three big circles is the two-dimensional probability simplex. Drawing consistent two-member Monte Carlo ensembles can be interpreted as randomly drawing points from this area. Placing an upper bound on the second ensemble member is equivalent to placing upper bounds on both q_1 and q_2. This constraint yields average values of q_1 and q_2 that are smaller than $1/3$. Even though there is no lower bound greater than zero on q_3, the average value of q_3 is larger than $1/3$ when sampling only from the gray area as opposed to sampling from the complete simplex.

The above conclusions are in agreement with the rank histograms of Fig. 1. Stratification along the median leads to a single step in the middle of the rank histogram. For the IQR and range, the criterion depends on two ensemble members. In both histograms we observe steps in the respective positions and plateaus in between the steps. If we stratify along mean and standard deviation, all ensemble members must be considered in order to calculate the criterion. Hence, there is a step between every two histogram bars, leading to a pattern without plateaus.

In appendix A we use the properties of the Dirichlet distribution to prove in full generality that stratification along the kth largest ensemble member yields the following steplike pattern:

$$\mathbb{E}[q_i \,|\, e_{[k]} < \tilde{\tau}] = c_1 < \frac{1}{K+1} \quad \forall i \leq k, \tag{18}$$

$$\mathbb{E}[q_i \,|\, e_{[k]} < \tilde{\tau}] = c_2 > \frac{1}{K+1} \quad \forall i > k, \tag{19}$$

which fits exactly the median pattern in Fig. 1 under the low stratum. The important insight is that under stratification

along a single ensemble member, the rank histogram will have two plateaus of constant heights c_1 and c_2. The expected heights of the bars in each individual plateau are exactly equal. Equations (18) and (19) provide the most complete description of the effect of stratification along the kth ensemble member if the forecast distribution is unknown. The constants c_1 and c_2 can be known precisely only if the forecast distribution is known at each instance.

By assuming that $\mathbb{P}(e_{[k]} < \tilde{\tau}) = 1/2$ (i.e., by assuming that the threshold $\tilde{\tau}$ is chosen so as to partition the original collection of ensembles into two strata of equal size), the expected rank probability in the high stratum is given by

$$\mathbb{E}[q_i \,|\, e_{[k]} > \tilde{\tau}] = \frac{2}{K+1} - c_1 > \frac{1}{K+1} \quad \forall i \le k, \tag{20}$$

$$\mathbb{E}[q_i \,|\, e_{[k]} > \tilde{\tau}] = \frac{2}{K+1} - c_2 < \frac{1}{K+1} \quad \forall i > k. \tag{21}$$

This is the same as saying that the sum of rank histograms of the strata have to add up to the rank histogram of the original ensemble, whose expected height is given by $1/(K+1)$.

6. Testing for ensemble consistency under stratification

A forecaster who wants to assess the statistical consistency of a forecast ensemble by stratifying along functions of the ensemble members has to account for the emergence of nonflat rank histograms even if the original ensemble is perfectly consistent. In this section we present a possible method to this end. We derive a χ^2 test for ensemble consistency under a known stratification pattern.

Let us first recall a standard statistical test for uniformity of the verification rank distribution. If a rank histogram is constructed from a finite number of samples, the bar heights are subject to random fluctuations and the histogram is never completely flat. The significance of deviations from flatness under such fluctuations can be assessed by means of the Pearson χ^2 test (Pearson 1900). Different tests for flatness of rank histograms exist (e.g., Elmore 2005; Bröcker 2008; Jolliffe and Primo 2008). Consider the case where we have N forecast–verification pairs of a K-member ensemble. The actually observed bar heights of the corresponding rank histogram are denoted by o_j, where $j = 1, \ldots, J$, and $J = K + 1$. The expected bar heights under the null hypothesis of statistical consistency are given by $\mathbb{E}[o_j] =: o^* = N/J$. Under the assumption that the null hypothesis is true, the test statistic

$$X^2 = \sum_{j=1}^{J} \frac{(o_j - o^*)^2}{o^*} \tag{22}$$

has a χ^2 distribution with $J - 1$ degrees of freedom. The p value of X^2 under this distribution can be used to conduct a hypothesis test at a certain confidence level. All histograms in Fig. 1 yield p values that are essentially equal to zero, indicating that uniformity of the verification rank distribution can be rejected at very high confidence levels. In contrast, the histogram that results from summing the low- and high-stratum histograms yields a p value of 0.26, thus substantiating statistical consistency of the original ensemble.

We have concluded in section 4 that stratification along functions of the ensemble alters the null hypothesis of equal bar heights of the rank histogram. The shape of the pattern that is introduced can be inferred if the stratification criterion is known. We have concluded that, if stratification is applied along the kth largest ensemble member $e_{[k]}$ then the stratification pattern of the rank histogram can be described by the J-dimensional vector:

$$\mathbf{c}(\theta) = \mathbf{c}_0 + \Bigg(\underbrace{\theta}_{k \text{ times}}, \underbrace{-\frac{k}{J-k}\theta}_{J-k \text{ times}} \Bigg), \tag{23}$$

where $\mathbf{c}_0 = (1/J, \ldots, 1/J)$ is the null hypothesis without stratification, $J = K + 1$ is the number of bars in the rank histogram, and the second term of the rhs of Eq. (23) is a vector whose elements sum to zero. The parameter θ indicates the strength of the pattern induced by stratification.

In appendix A we show that under such a pattern as the new null hypothesis, a χ^2 test can be formulated. Consider an observed rank histogram with the height of the jth bar given by o_j, that is rank j has occurred o_j times. We then further denote by $N_{(m,n)} = \sum_{j=m+1}^{n} o_j$, the number of instances at which the verification rank is in the interval $(m, n]$, and $N = N_{(0,J)}$, the total number of instances. Then the test statistic:

$$-2\left(N_{(0,k)} \log \frac{N_{(0,k)}}{Nk} + N_{(k,J)} \log \frac{N_{(k,J)}}{N(J-k)} - \sum_{j=1}^{J} o_j \log \frac{o_j}{N} \right), \tag{24}$$

has a χ^2 distribution with $J - 2$ degrees of freedom if the rank histogram really assumes the pattern given by Eq. (23). The dependence on the unknown parameter θ is eliminated because the test makes use only of its maximum-likelihood estimate obtained from the observed rank histogram. This is the reason why the χ^2 distribution has only $J - 2$ degrees of freedom. In appendix A a second χ^2 test is given for the case when the rank histogram pattern has two steps (as in the case of stratification along a function of two ensemble members, such as the IQR or the range).

Unfortunately, our reasoning cannot be extended to cases where the stratification criterion depends on all ensemble members, such as the ensemble mean or standard deviation, which have been employed in previous studies (Hamill and Colucci 1997, 1998). We argue that median and IQR, to which our theory applies, are suitable alternatives to the mean and standard deviation, respectively.

The derivation of the test statistic given in Eq. (24) makes use of the maximum-likelihood estimate of the free parameter θ of Eq. (23). This estimate is given by

$$\hat{\theta} := \frac{N_{(0,k)}}{Nk} - \frac{1}{J}. \tag{25}$$

The value of $\hat{\theta}$ under different perfect ensemble scenarios provides further insight. It indicates how strongly the rank histogram of the ensemble is affected by a certain kind of stratification. The $\hat{\theta}$ approaches zero as the rank histogram becomes flat. We consider sampling with equal probability from two distinct Gaussian distributions. Let one distribution have zero mean and take the mean μ of the other distribution as a parameter. Both distributions have unit variance. An ensemble forecast–verification dataset is produced by sampling $N = 10^5$ consistent ensembles randomly from each distribution. The combined dataset is then stratified along the ensemble median $e_{[(K+1)/2]}$. Depending on the ensemble size K and the parameter μ, stratification patterns with different values of $\hat{\theta}$ are introduced.

We take $J\hat{\theta} \in [-1, 1]$ as a measure of the strength of the stratification pattern. In Fig. 3 we plot $J\hat{\theta}$ for the high- and low-median stratum and for different ensemble sizes K as a function of μ. One can observe that the strength of the pattern decreases with increasing ensemble size K: while the maximum value of $J\hat{\theta}$ is close to 0.25 for $K = 11$, it barely exceeds 0.1 if $K = 51$. The $J\hat{\theta}$ is largest around $\mu \approx 0$ where the two distributions from which the ensemble–verification pairs are sampled are almost identical. For $\mu = 0$ all fluctuations of the median are caused by fluctuations due to the finiteness of the ensemble. On the other hand, $J\hat{\theta}$ approaches zero at values of $\mu = \pm 2$, indicating that at these values, stratification hardly induces any pattern.

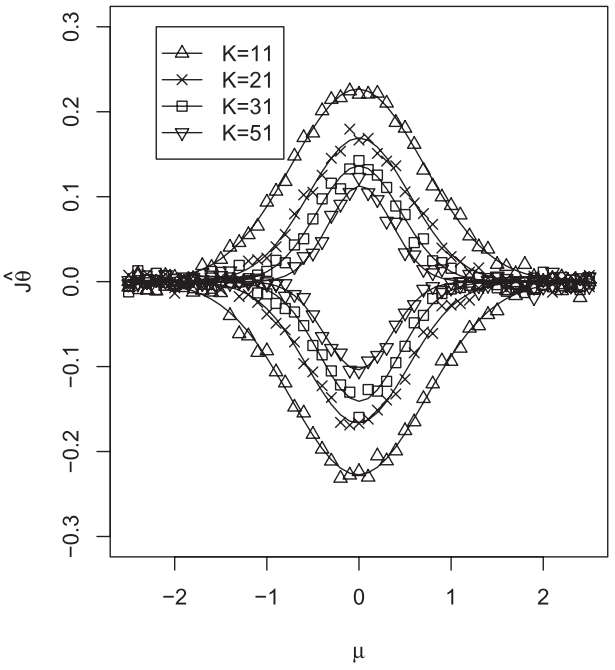


FIG. 3. Ensemble–verification pairs are drawn with equal probability from either a standard Gaussian or a Gaussian with mean μ and variance 1. The collection of ensembles is subsequently stratified into two strata along the ensemble median. The relative strength of the pattern Eq. (23) [using $k = (K + 1)/2$] is given by $J\hat{\theta}$. It is plotted as a function of μ. Under this pattern, the positive values of $J\hat{\theta}$ correspond to the high-median stratum and the negative ones to the low-median stratum. Gaussian curves have been fitted as a guide to the eye. The strength of the pattern decreases if the two distributions are well separated (i.e., for $|\mu| > 2$). Larger ensembles are less sensitive to stratification patterns.

7. Avoiding stratification patterns

In this section we present possible ways to avoid the emergence of patterns under stratification altogether. First of all let us recall the mechanism that leads to nonuniform values of $\mathbb{E}[\mathbf{q}]$. From the discussion of section 5 we know that it is the bound on some of the q_i that is introduced by thresholding certain ensemble members by $\tilde{\tau}$ [see Eq. (17)]. More specifically, it is $F(\tilde{\tau})$, denoted by τ, which restricts the q_i. If one can assure, according to Eq. (17), that τ is very close to 1 in the low stratum and very close to 0 in the high stratum, the patterns introduced by stratification vanish because the lower and upper bounds on the rank probabilities q_i are then asymptotically equal to 0 and 1, respectively.

As an illustrative example consider the case where the forecast distribution shows a regimelike behavior. Such behavior is given for example for large values of μ in the scenario under which Fig. 3 was constructed. If the median is collected from each of the N ensembles, distinct values of the median around 0 on the one hand, and around $\mu \gg 0$ on the other hand will occur. If the threshold that groups the instances into their strata is set to $\mu/2$, the two distributions are well separated by stratification. If the observed median is in the high stratum, the ensemble is almost certainly sampled from the Gaussian with positive mean. The cdf from which the ensemble was drawn, evaluated at $\tilde{\tau} = \mu/2$ is close to zero under this distribution. Thus, the lower bound on

the q_i is likewise close to zero such that the sampling area on the simplex will not be significantly diminished. Similarly, the ensemble cdf evaluated at $\tilde{\tau}$ in the low stratum will be very close to one. The area from which the vectors $\mathbf{q}$ are sampled then hardly differs from the full probability simplex in either of the strata. If such regimelike behavior of the strata can be assumed, stratification does indeed lead to asymptotically flat rank histograms under all strata. However, note that under the same scenario, stratifying along the IQR, for example, can still lead to artifacts since this quantity does not behave regimelike.

From this discussion, it becomes apparent that stratification should be physically justified. There should be upfront evidence for well-separated strata. If the grouping of ensemble–verification pairs into the different strata is influenced by statistical fluctuations because of the finiteness of the ensemble, which is the case if $\mu \approx 0$ in the above scenario, artifacts are likely to occur.

If it is not clear that stratification based on the ensemble can clearly distinguish between truly different regimes like in the example above, caution must be exercised. In a recent correspondence about conditional exceedance probabilities (see Mason et al. 2007; Bröcker et al. 2011; Mason et al. 2011), a similar effect as the one elaborated in this paper was discussed. Mason et al. (2011) propose a method to avoid artifacts that result from conditioning a consistency analysis of a MCE on the ensemble itself. They consider splitting the K-member ensemble at each instance into two daughter ensembles, by sampling randomly without replacement $K/2$ times from the full ensemble. One daughter ensemble is exclusively used to calculate the stratification criterion (say a certain quantile of the ensemble or its spread) and subsequently discarded. Only the second daughter ensemble is then actually stratified, using the value of the criterion obtained from the first daughter ensemble, and subject to evaluation by, for example, the rank histogram. This procedure renders the evaluated ensemble independent from the calculation of the stratum. Constraints on the ensemble members as in Eq. (17) then disappear and artifacts are avoided.

This method works not only if the stratified ensembles are evaluated by means of rank histograms but for any consistency and reliability analysis. The downside of this approach is that the ensemble that is actually evaluated is smaller and possibly contains less forecast information than the original ensemble. The rank histogram would only contain half the original number of bins and important details might get blurred by this coarsening. Notwithstanding the above, we consider such an approach a versatile alternative to the method presented in section 6.

8. Discussion and conclusions

We have considered the effect of stratifying statistically consistent Monte Carlo ensembles (MCEs) along functions of the ensemble. We focused mainly on the special case of stratification along a single member of the ordered ensemble. We concluded that ensembles that are the result of stratifying a consistent MCE are themselves not necessarily consistent, which leads to their rank histograms not being flat. We have provided arguments based on a perfect ensemble simulation study and based on a mathematical formalization of MCEs. The latter formalization has lead to a statistical test for the consistency of stratified MCEs.

In section 2, the Dirichlet distribution was shown to be of relevance to ensemble forecasts. In the example given here, the resulting Dirichlet distribution turned out to be the trivial version that amounts to uniform sampling. Notwithstanding this, the properties of the Dirichlet distribution were useful in obtaining rigorous results about the artifacts introduced into the rank histogram due to stratification. We expect the formalism developed in this article to be relevant to further questions regarding forecast ensembles.

In section 3 we have discussed the possible benefits of ensemble stratification. A possible direction of future research would be considering not only rank histograms of stratified ensembles but also skill scores, reliability diagrams, and ROC plots. We expect stratification to lead to similar artifacts in such analyses.

The perfect ensemble simulations of section 4 are a simple framework to test assumptions about forecast ensembles. We propose that new ensemble evaluation techniques should be tested using such a perfect ensemble in order to assure that the technique produces results under the null hypothesis of a consistent ensemble that are in agreement with what the forecaster expects. Ensemble stratification provides an example where this is not necessarily the case.

Section 5 provides a formal description of the effect of ensemble stratification on the rank histogram. The rank histogram pattern of a consistent MCE stratified along the kth largest member could be derived rigorously. However, we point out again that no rigorous derivations of the stratification patterns were presented for the other criteria, including the IQR and the range. However, we have applied the statistical test given by Eq. (C10) in many different perfect ensemble simulations using different ensemble sizes, different distributions from which ensembles and verifications were drawn, and different criteria that involve two members of the ordered ensemble. Based on the fact that all these tests produced consistent results and based on the handwaving arguments about such strata in section 5 we hypothesize that stratification along

the difference between the mth and kth largest ensemble member indeed produces the pattern given by Eq. (C9).

An advantage of stratification along median, IQR and range is that these criteria introduce highly artificial, steplike patterns. For once, these patterns have been shown to be better controllable in the χ^2 test than the smooth patterns introduced by mean and standard deviation. On the other hand, these patterns would probably not be observed as a consequence of a real ensemble inconsistency, such as underdispersiveness or bias, where more smooth patterns would be expected. Thus, these quantile-based criteria are both better controllable and better distinguishable from patterns arising through genuine ensemble deficiencies. Consequently, the proposed χ^2 test is also less likely to take a true pattern as an artifact of stratification, thus failing to detect this true inconsistency. We therefore strongly advocate stratification along the quantile-based criteria as opposed to mean and standard deviation.

The stratification criteria presented and discussed in this paper are not the only ones possible. Stratification along season, along indices describing large-scale atmospheric behavior (such as the ENSO index) or even along the verification can in principle be applied. These criteria are external to the ensemble (i.e., they do not depend on the ensemble directly) in contrast to the criteria we considered in the present paper. However, caution must be exercised nonetheless. Stratification along the verification would very likely induce stratification patterns in a consistent ensemble. If the verification is, say, in the larger-than-average stratum, it is also more likely to occupy one of the higher ranks in the forecast ensemble, thus leading to a sloped rank histogram. Furthermore, the value of stratification along the verification is questionable, since the verification is not known at forecast time. Hence, no ensemble correction could possibly be applied based on knowledge about different ensemble deficiencies under different strata of the verification. The current meteorological regime, for example, might be inferred from the ensemble, and might thus not be entirely independent from the ensemble either. For this reason, stratification along the current regime might lead to stratification artifacts as well. We reserve a more detailed analysis of such criteria for future studies.

Acknowledgments. The authors acknowledge helpful comments from two anonymous referees.

APPENDIX A

The Dirichlet Distribution

In this section we review the Dirichlet distribution and list some of its properties which are relevant to the discussions in this paper. The reader is referred to Frigyik et al. (2010) for an excellent introduction to the Dirichlet distribution and related processes. Introductory texts on Bayesian analysis usually contain material on the Dirichlet distribution (e.g., Bernardo and Smith 1994) because of its role as a conjugate prior for the multinomial distribution.

A probability mass function (pmf) of length J defines a discrete probability distribution over J categories. The $(J-1)$-dimensional probability simplex is the set of J-dimensional vectors whose elements are nonnegative and sum to one. It is a $(J-1)$-dimensional surface embedded in J dimensions. Every vector $\mathbf{q} = (q_1, \ldots, q_J)$ that lies on the $(J-1)$-dimensional probability simplex can be interpreted as a J-dimensional pmf.

The Dirichlet distribution is a distribution over the probability simplex. It models randomly drawn pmfs. Let $\tilde{\mathbf{q}} = (q_1, \ldots, q_{J-1})$ be a $(J-1)$-dimensional random vector that satisfies $0 < q_j < 1 \; \forall j$ and $\sum_{j=1}^{J-1} q_j < 1$. Furthermore, let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)$ be a J-dimensional vector whose elements satisfy $\alpha_j \geq 0 \; \forall j$. Then the J-dimensional vector $\mathbf{q} = (q_1, \ldots, q_{J-1}, 1 - \sum_{j=1}^{J-1} q_j)$ has a Dirichlet distribution with parameters $\boldsymbol{\alpha}$ if its probability density is given by

$$p(\mathbf{q} \mid \boldsymbol{\alpha}) = \gamma \left(\prod_{j=1}^{J-1} q_j^{\alpha_j - 1} \right) \left(1 - \sum_{j=1}^{J-1} q_j \right)^{\alpha_J - 1}, \quad \text{(A1)}$$

where

$$\gamma = \frac{\Gamma\left(\sum_{j=1}^{J} \alpha_j \right)}{\prod_{j=1}^{J} \Gamma(\alpha_j)}, \quad \text{(A2)}$$

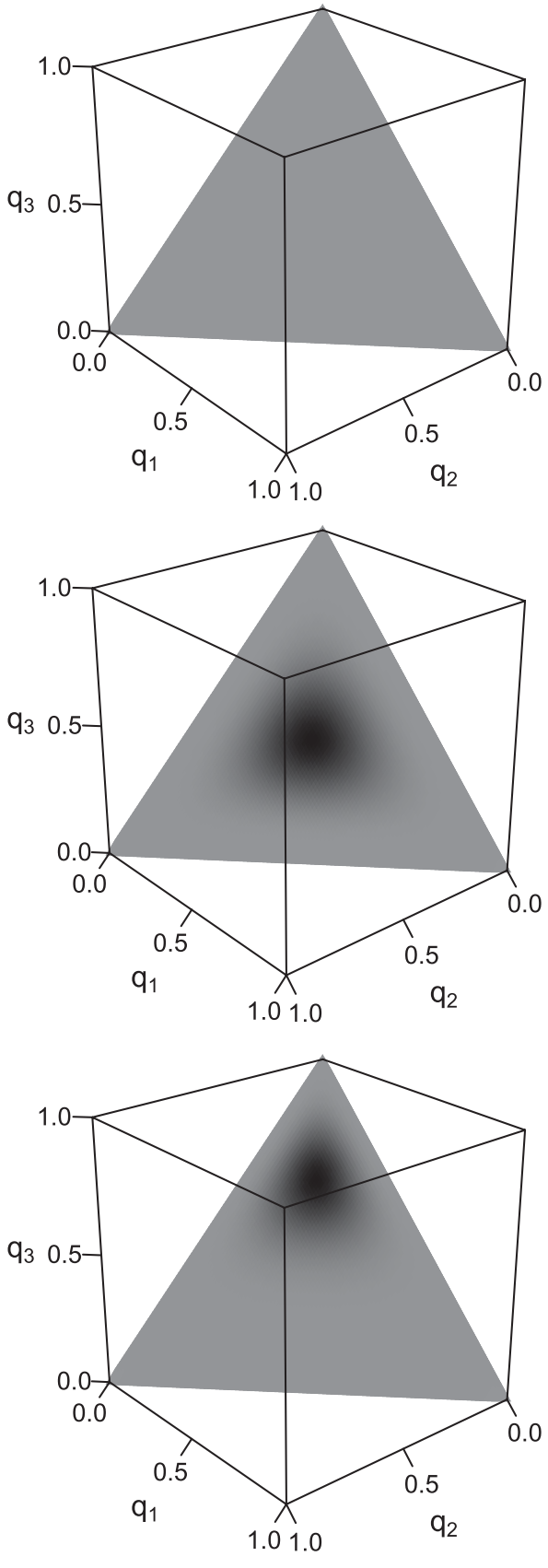
and $\Gamma(\cdot)$ is the Gamma function. If $\mathbf{q}$ is distributed according to Eq. (A1), we write $\mathbf{q} \sim \mathrm{Dir}(\boldsymbol{\alpha})$. For the case $J = 2$, the Dirichlet distribution reduces to the beta distribution with parameters α_1 and α_2.

A more convenient way to define the Dirichlet distribution is to set $p(\mathbf{q} \mid \boldsymbol{\alpha})$ to zero if $\mathbf{q}$ does not lie on the probability simplex and

$$p(\mathbf{q} \mid \boldsymbol{\alpha}) = \gamma \prod_{j=1}^{J} q_j^{\alpha_j - 1} \quad \text{(A3)}$$

otherwise.

The parameters $\boldsymbol{\alpha}$ determine the way in which the points $\mathbf{q}$ are sampled from the probability simplex. The case $\boldsymbol{\alpha} = (1, \ldots, 1)$ amounts to uniform sampling since then $p(\mathbf{q} \mid \boldsymbol{\alpha}) = \text{const}$. If $\alpha_j = c \; \forall j$ with $c < 1$, points are



sampled closer to the vertices of the simplex and if $c > 1$, sampling is more concentrated in the center of the simplex. Unequal α_j lead to a noncentral distribution with unequal expectation values of the $\mathbf{q}$ components (see Fig. A1).

If $\mathbf{q} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, the expectation of $\mathbf{q}$ is given by

$$\mathbb{E}[\mathbf{q}] = \frac{\boldsymbol{\alpha}}{\sum_j \alpha_j} \tag{A4}$$

(Frigyik et al. 2010). That means that for any Dirichlet distribution with parameters $\alpha_j = \mathrm{const}\ \forall j$, the expectation of q_j is equal to $1/J$. The marginal distribution of q_j is a beta distribution with parameters α_j and $\sum_{i \neq j} \alpha_i$ (Frigyik et al. 2010). If $\alpha_i = \alpha_j$, the components q_i and q_j are exchangeable, that is, the joint density of q_i and q_j satisfies

$$p_{\ldots,q_i,\ldots,q_j,\ldots}(\ldots,u,\ldots,v,\ldots) = p_{\ldots,q_i,\ldots,q_j,\ldots}(\ldots,v,\ldots,u,\ldots), \tag{A5}$$

which follows from the definition of the Dirichlet density in Eq. (A3). Consider the vector $\mathbf{Q} = (\sum_{i=1}^{k} q_i, q_{k+1}, \ldots, q_J)$. The aggregation property of the Dirichlet distribution states that if $\mathbf{q} \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_J)$ then (Frigyik et al. 2010)

$$\mathbf{Q} \sim \mathrm{Dir}\left(\sum_{i=1}^{k} \alpha_i, \alpha_{k+1}, \ldots, \alpha_J\right). \tag{A6}$$

APPENDIX B

Proof of the Step Pattern

Consider stratification along a single ensemble member $e_{[k]}$. Then in the low stratum the height of the jth bar of the rank histogram is proportional to $\mathbb{E}[q_j \,|\, e_{[k]} < \tilde{\tau}]$. Applying the PIT to the condition yields $\mathbb{E}[q_j \,|\, \sum_{i=1}^{k} q_k < \tau]$, where $\tau \leq 1$ is the ensemble cdf evaluated at $\tilde{\tau}$. Consider the vector $\mathbf{Q} = (\sum_{i=1}^{k} q_i, q_{k+1}, \ldots, q_J)$. By the aggregation property of the Dirichlet distribution we have $\mathbb{E}\mathbf{Q} = (k, 1, \ldots, 1)/J$. Conditioning $Q_1 < \tau$ yields $\mathbb{E}[Q_1 \,|\, Q_1 < \tau] < k/J$. Thus, since the elements of $\mathbb{E}\mathbf{Q}$ have to sum to one, at least one of the $Q_{i>1}$ must be

←

FIG. A1. Examples of Dirichlet densities on the two-dimensional probability simplex. Darker colors indicate higher densities. (top) $\boldsymbol{\alpha} = (1, 1, 1)$. The density is uniform over the simplex. (middle) $\boldsymbol{\alpha} = (5, 5, 5)$. Sampling is more concentrated in the center of the simplex. (bottom) $\boldsymbol{\alpha} = (4, 4, 10)$. Values of q_3 are sampled closer to one, while q_1 and q_2 are sampled closer to zero.

larger than $1/J$. The $Q_{i>1}$ are exchangeable because the corresponding α_i are equal. Exchangeability implies that their expectations must be equal. We conclude that $\mathbb{E}[Q_{i>1} \mid Q_1 < \tau] = \mathbb{E}[q_{i>k} \mid e_{[k]} < \tilde{\tau}] = \text{const} > 1/J$. From exchangeability it follows that $\mathbb{E}[q_{i\leq k} \mid e_{[k]} < \tilde{\tau}] = \text{const} < 1/J$. The high stratum pattern can be inferred from the following relation:

$$\mathbb{E}q_i = \mathbb{E}[q_i \mid e_{[k]} < \tilde{\tau}]\mathbb{P}(e_{[k]} < \tilde{\tau}) + \mathbb{E}[q_i \mid e_{[k]} > \tilde{\tau}]\mathbb{P}(e_{[k]} > \tilde{\tau}) \tag{B1}$$

$$= \frac{1}{2}(\mathbb{E}[q_i \mid e_{[k]} < \tilde{\tau}] + \mathbb{E}[q_i \mid e_{[k]} > \tilde{\tau}]), \tag{B2}$$

where we assume that $\mathbb{P}(e_{[k]} < \tilde{\tau}) = \mathbb{P}(e_{[k]} > \tilde{\tau}) = 1/2$, that is each stratum contains exactly one-half of all cases. We conclude that the pattern in the high stratum is the reversed version of the pattern derived for the low stratum. If the expectation decreases in the low stratum, it has to increase in the high stratum according to Eq. (B2).

APPENDIX C

Derivation of Eq. (24)

A theorem by Wilks (1938) makes a statement about the asymptotic distribution of generalized likelihood ratios (see also Mood et al. 1974). Assume two parameter spaces Θ and $\Theta_0 \subset \Theta$ whose elements parameterize candidate distributions that have generated the data points $\mathbf{y} = (y_1, \ldots, y_N)$. Let the codimension of Θ_0 in Θ be equal to a and let $p(y_t \mid \theta)$ be the likelihood of the tth data point, given that the distribution is parameterized by θ. If the data $\mathbf{y}$ were indeed generated by a distribution parameterized by a $\theta \in \Theta_0$ then

$$-2\log\frac{\sup_{\theta\in\Theta_0}\prod_{t=1}^{N} p(y_t \mid \theta)}{\sup_{\theta\in\Theta}\prod_{t=1}^{N} p(y_t \mid \theta)} \sim \chi^2_a, \tag{C1}$$

that is, the ratio between the maximized likelihoods of the data in both parameter spaces, transformed by $-2\log(\cdot)$, has a χ^2 distribution with a degrees of freedom in the limit $N \to \infty$.

In an unstratified, consistent K-member ensemble the expectation of its J-dimensional rank distribution $\mathbf{q}_0$ is given by

$$\mathbb{E}[\mathbf{q}_0] = \left(\frac{1}{J}, \ldots, \frac{1}{J}\right) =: \mathbf{c}_0, \tag{C2}$$

which yields a flat rank histogram. Stratification along a function of the ensemble members leads to a pattern that we describe by a vector $\mathbf{c}(\theta), \theta \in \Theta_0$, whose elements sum to zero. The new expected rank distribution is given by the following superposition:

$$\mathbb{E}[\mathbf{q}] = \mathbf{c}_0 + \mathbf{c}(\theta). \tag{C3}$$

The elements of $\mathbb{E}[\mathbf{q}]$ in Eq. (C3) sum to one because the elements of $\mathbf{c}(\theta)$ sum to zero.

We apply the Wilks theorem [Eq. (C1)] to formulate a hypothesis test for rank histograms that are the result of such a process. Let the data point $y_t = j$ if verification rank j occurs on the tth instance. Then we have for the height of the jth bar of the rank histogram $o_j = \sum_{t=1}^{N} \mathbb{I}(y_t = j)$, where $\mathbb{I}(\cdot)$ is the indicator function.

We take the $(J-1)$-dimensional probability simplex as the parameter space Θ, which contains all possible pmfs $\mathbb{E}[\mathbf{q}]$ (including the nonflat ones) that could have lead to the observed rank histogram with bar heights o_j. The maximum likelihood parameter $\mathbf{q} \in \Theta$ for the data $\mathbf{y}$ is given by $q_j = o_j/N$, which follows from setting the derivative of the log-likelihood with respect to q_j to zero and solving for q_j. The logarithm of the denominator of Eq. (C1) is then given by

$$\log\sup_{\mathbf{q}\in\Theta}\prod_{t=1}^{N} p(y_t \mid \mathbf{q}) = \sum_{j=1}^{J} o_j \log\frac{o_j}{N}. \tag{C4}$$

To illustrate the calculation of the numerator of Eq. (C1), we return to the simple example of stratification along the single ensemble member $e_{[k]}$. From the discussion of section 5 and the proof of appendix A we know that the ensuing pattern $\mathbf{c}(\theta)$ must have the form given by Eq. (23). We denote the union of all possible $\mathbf{c}(\theta)$ as Θ_0. Since every element of Θ_0 is a pmf, we have $\Theta_0 \subset \Theta$. Furthermore the dimension of Θ_0 is 1 since it is parameterized by a single parameter θ. It follows that the codimension of Θ_0 in Θ is equal to $J-2$. The likelihood of the tth datum y_t as a function of θ is given by

$$p(y_t \mid \theta) = \frac{1}{J} + c_{y_t}(\theta). \tag{C5}$$

Denote $N_{(m,n)} = \sum_{j=m+1}^{n} o_j$, the number of instances where the verification rank $y_t \in (m, n]$. By setting the derivative of the log-likelihood with respect to θ to zero and solving for θ we obtain $\hat{\theta}$, the maximum likelihood estimator of θ, which is given by

$$\hat{\theta} := \operatorname*{argmax}_{\theta\in\Theta_0}\prod_{t=1}^{N} p(y_t \mid \theta) = \frac{N_{(0,k)}}{Nk} - \frac{1}{J}. \tag{C6}$$

Thus, we get for the logarithm of the numerator of Eq. (C1):

$$\log \sup_{\theta \in \Theta_0} \prod_{t=1}^{N} p(y_t \mid \theta) = N_{(0,k)} \log \frac{N_{(0,k)}}{Nk} + N_{(k,J)} \log \frac{N_{(k,J)}}{N(J-k)}. \tag{C7}$$

Substituting Eqs. (C4) and (C7) into Eq. (C1), the generalized likelihood ratio test reads

$$-2\left(N_{(0,k)} \log \frac{N_{(0,k)}}{Nk} + N_{(k,J)} \log \frac{N_{(k,J)}}{N(J-k)} - \sum_{j=1}^{J} o_j \log \frac{o_j}{N}\right) \sim \chi^2_{J-2}. \tag{C8}$$

If the stratum is defined by the difference between two ensemble members $e_{[m]}$ and $e_{[k]}$ ($m > k$), which is the case in stratification along the range or IQR, we hypothesize a pattern of the following form:

$$\mathbf{c}(\theta_1, \theta_2, \theta_3) = \mathbf{c}_0 + (\underbrace{\theta_1}_{k \text{ times}}, \underbrace{\theta_2}_{m-k \text{ times}}, \underbrace{\theta_3}_{J-m \text{ times}}). \tag{C9}$$

Using similar arguments as in the proof of Eq. (C8) one can show that such a pattern leads to the hypothesis test:

$$-2\left(N_{(0,k)} \log \frac{N_{(0,k)}}{Nk} + N_{(k,m)} \log \frac{N_{(k,m)}}{N(m-k)} + N_{(m,J)} \log \frac{N_{(m,J)}}{N(J-m)} - \sum_{j=1}^{J} o_j \log \frac{o_j}{N}\right) \sim \chi^2_{J-3}, \tag{C10}$$

where the χ^2 distribution now has $J - 3$ degrees of freedom because Θ_0 is two-dimensional.

Note that the above theory does not apply when a stratification criterion depends on all K ensemble members. If this is the case, we would assume K steps in the rank histogram, which requires K parameters to describe the pattern. Thus, Θ would be equal to Θ_0 and their codimension is zero. The test in Eq. (C1) is then not defined.

REFERENCES

Anderson, J., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530.

Bernardo, J., and A. Smith, 1994: *Bayesian Theory*. Wiley Series in Probability and Mathematical Statistics, Wiley, 586 pp.

Bröcker, J., 2008: On reliability analysis of multi-categorical forecasts. *Nonlinear Processes Geophys.,* **15,** 661–673.

——, S. Siegert, and H. Kantz, 2011: Comments on ''Conditional exceedance probabilities.'' *Mon. Wea. Rev.,* **139,** 3322–3324.

Elmore, K., 2005: Alternatives to the chi-square test for evaluating rank histograms from ensemble forecasts. *Wea. Forecasting,* **20,** 789–795.

Frigyik, B., A. Kapila, and M. Gupta, 2010: Introduction to the Dirichlet distribution and related processes. Dept. of Electrical Engineering, University of Washington, Tech. Rep. 6, 27 pp.

Hamill, T., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.,* **129,** 550–560.

——, and S. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327.

——, and ——, 1998: Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.

Hyndman, R., and Y. Fan, 1996: Sample quantiles in statistical packages. *Amer. Stat.,* **50** (4), 361–365.

Jolliffe, I., and C. Primo, 2008: Evaluating rank histograms using decompositions of the chi-square test statistic. *Mon. Wea. Rev.,* **136,** 2133–2139.

Mason, S., J. Galpin, L. Goddard, N. Graham, and B. Rajartnam, 2007: Conditional exceedance probabilities. *Mon. Wea. Rev.,* **135,** 363–372.

——, M. Tippett, A. Weigel, L. Goddard, and B. Rajaratnam, 2011: Reply. *Mon. Wea. Rev.,* **139,** 3325–3327.

Mood, A., F. Graybill, and D. Boes, 1974: *Introduction to the Theory of Statistics.* 3rd ed. McGraw-Hill, 480 pp.

Mullen, S., and R. Buizza, 2002: The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Wea. Forecasting,* **17,** 173–191.

Pearson, K., 1900: X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag. Series 5,* **50** (302), 157–175.

Peel, S., and L. Wilson, 2008: A diagnostic verification of the precipitation forecasts produced by the Canadian ensemble prediction system. *Wea. Forecasting,* **23,** 596–616.

Siegert, S., J. Bröcker, and H. Kantz, 2011: Predicting outliers in ensemble forecasts. *Quart. J. Roy. Meteor. Soc.,* **137** (660), 1887–1897.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. ECMWF Workshop on Predictability,* Reading, United Kingdom, European Centre for Medium-Range Weather Forecasts, 1–25.

Wilks, S., 1938: The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.,* **9,** 60–62.