On the Reliability of the Rank Histogram

D. S. Wilks

Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York

(Manuscript received 5 April 2010, in final form 7 June 2010)

ABSTRACT

Ensemble consistency is a name for the condition that an observation being forecast by a dynamical ensemble is statistically indistinguishable from the ensemble members. This statistical indistinguishability condition is meaningful only in a multivariate sense. That is, it pertains to the joint distribution of the ensemble members and the observation. The rank histogram has been designed to assess overall ensemble consistency, but mistakenly employing it to assess only restricted aspects of this joint distribution (e.g., the climatological distribution) leads to the incorrect conclusion that the verification rank histogram is not a useful diagnostic for good behavior of ensemble forecasts. The potential confusion is analyzed in the context of an idealized multivariate Gaussian model of forecast ensembles and their corresponding observations, and it is shown that the rank histogram does correctly assess the consistency of forecast ensembles.

1. Introduction

The verification rank histogram (RH; Anderson 1996; Hamill and Colucci 1997; Talagrand et al. 1997) is a commonly used diagnostic for ensemble forecasts of scalar predictands. For each of a collection of $n_{\text{ens}}$ member forecast ensembles, the ensemble members and the verifying observation are jointly ranked in ascending order. The rank of the observation within each group of $n_{\text{ens}} + 1$ values is tabulated over a sample of $n$ forecasts and observations, and a histogram is constructed from these ranks. The ideal result is that the RH is (apart from sampling variations) flat, or uniform, which would be suggestive of the ensemble prediction system performing well with respect to representing the forecast uncertainty.

The idea behind the RH and its ideal shape being flat is the notion of a “consistent” (Anderson 1997; Wilks 2006b) or “perfect” (Buizza 1997; Johnson and Bowler 2009; Palmer et al. 2006) ensemble, for which the verifying observation is statistically indistinguishable from the ensemble of forecasts, again as evaluated over a large number of forecast cases. Ensemble consistency is sometimes also called ensemble “reliability” (e.g., Toth et al. 2003), although ensemble consistency does not necessarily imply that probability forecasts constructed from the ensemble are reliable in the sense of conditional outcome relative frequencies being equal to the forecast probabilities (Murphy 1973; Murphy and Winkler 1987), yielding a 45° calibration function on a reliability diagram, unless either the ensemble size is relatively large or the forecasts are reasonably skillful, or both (Richardson 2001).

Another diagnostic that is often used to evaluate ensemble under- or overdispersion in aggregate is comparison of the average (again, over $n$ forecast cases) ensemble standard deviation to the RMSE of the ensemble mean forecast (Buizza 1997; Johnson and Bowler 2009; Palmer et al. 2006; Toth et al. 2003). Consistent ensembles should exhibit equality of these two quantities. Under-dispersed ensembles are expected to have average ensemble standard deviations that are smaller than the
ensemble-mean RMSE, and vice versa for overdispersed ensembles. However, unconditional forecast biases will inflate RMSE without affecting the ensemble dispersion (because ensemble dispersion is computed relative to the sample ensemble mean), so that this diagnostic cannot distinguish forecast bias from ensemble underdispersion.

It is not always understood that ensemble consistency must be interpreted in a multivariate sense. That is, it is necessary but not sufficient for the univariate ("climatological") distributions of the ensemble members and of the predictand to be the same. The residual-quantile-quantile (R-Q-Q) plot (Marzban et al. 2011) is a useful tool for diagnosing this aspect of ensemble performance. Ensemble consistency further requires that the joint distribution of the ensemble members will be unchanged when the verifying observation is substituted for any one of the ensemble members. Thus, if the ensemble members and their verifying observations have Gaussian climatological distributions, then ensemble consistency requires not only that these distributions have the same means and variances (i.e., the same climatologies), but also that the correlation between pairs of ensemble members is equal to the correlation between an ensemble member and the verifying observation.

Recently Marzban et al. (2010) analyzed the behavior of the RH, incorrectly employing it to assess climatological dispersion. They concluded that RHs are "not interpretable," claiming that the characteristic "U" and "mound" RH shapes cannot be used to diagnose ensemble under- and overdispersion, respectively. Using the same multivariate Gaussian model, this paper demonstrates that conclusion to be incorrect, and that it derives from an improper expectation that the RH should assess climatological dispersion. In a follow-up paper, Marzban et al. (2011) retract their initial claims, distinguishing between two components of reliability as proposed by Johnson and Bowler (2009), and advocating use of R-Q-Q plots to evaluate consistency of the model climatological distribution.

Section 2 outlines the multivariate Gaussian statistical model for the ensembles and their verifying observations introduced by Marzban et al. (2010, 2011). Section 3 presents a revised analysis of the Marzban et al. (2010) results, and section 4 provide the conclusions.

2. Multivariate Gaussian model of forecast ensembles and scalar observations

An instructive multivariate Gaussian model for the distribution of ensemble members and verifying observations is presented by Marzban et al. (2010, 2011). Within this multivariate model, the univariate (climatological) distribution for the observations is standard (i.e., zero mean and unit variance) Gaussian:

\[ y_0 \sim N(0,1), \]  
and the univariate distribution for each ensemble member is Gaussian with common mean \( \mu \) and (climatological) variance \( \sigma^2 \):

\[ y_i \sim N(\mu, \sigma^2), \quad i = 1, \ldots, n_{\text{ens}}. \]  

These independent univariate distributions for the observation and ensemble members composed the full statistical model of ensemble forecast behavior used by Hamill (2001). Equation (1) indicates in effect that the verifying observations are expressed as standardized anomalies. Consistent ensembles will exhibit \( \mu = 0 \) and \( \sigma^2 = 1 \), but these conditions are not sufficient to ensure a consistent ensemble when the ensemble members and the observation are not independent (which is the usual situation). Ensembles for which \( \mu \neq 0 \) will exhibit unconditional biases, but \( \sigma^2 = 1 \) does not imply that the ensembles are necessarily free from over- or underdispersion, in the sense of the observation being outside the ensemble too rarely or too frequently, as will be illustrated in section 3. For convenience, \( n_{\text{ens}} = 10 \) was used in Marzban et al. (2010) and will also be used here, although the interpretations do not depend on this choice.

The joint multivariate Gaussian distribution for the ensemble members and the verifying observation is defined by the means as specified in Eqs. (1) and (2), and the joint covariance matrix:

\[
\Sigma = \begin{bmatrix}
1 & R \sigma & R \sigma & \cdots & R \sigma \\
R \sigma & \sigma^2 & r \sigma^2 & \cdots & r \sigma^2 \\
R \sigma & r \sigma^2 & \sigma^2 & \cdots & r \sigma^2 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
R \sigma & r \sigma^2 & r \sigma^2 & \cdots & \sigma^2 
\end{bmatrix}. \]  

The variances on the diagonal of Eq. (3) are as specified in Eqs. (1) and (2). The correlation \( R \) is related to the typical accuracy of each ensemble member's forecast (e.g., Murphy 1988), so that \( R \) approaches 1 for nearly perfect predictability (if, also, the univariate distributions are the same so that \( \mu = 0 \) and \( \sigma^2 = 1 \)), and \( R \) approaches 0 for completely useless forecasts that are statistically independent of the verifying observations (zero correlation implies statistical independence for Gaussian variables). The correlation \( r \) characterizes the typical degree of similarity of the ensemble members to each other or the compactness of the ensemble. For
r approaching 1, the ensemble members become nearly identical to one another. For r = 0, each ensemble member is an independent random draw from its climatological distribution [Eq. (2)].

For a consistent ensemble, the compactness of the ensemble (forecast sharpness) must be commensurate with the accuracy of the ensemble members’ forecasts. If the individual member’s forecasts are highly accurate (R near 1, assuming μ = 0 and σ² = 1), each ensemble member must necessarily be very similar to all the others (similarly large r); otherwise, some ensemble members could be very different from the verifying observation, which would be inconsistent with R being near 1. Conversely, if the ensemble members are random draws from their common climatological distribution (r = 0), they cannot also yield accurate forecasts except by chance, implying small R also. Indeed, r ≪ R is inconsistent with the multivariate Gaussian covariance matrix in Eq. (3), which must be positive definite. For the verifying observation to be statistically indistinguishable from the ensemble members in the multivariate sense, exchanging it for an ensemble member must not change the joint covariance matrix for the ensemble members [lower-right quadrant of the covariance matrix Σ in Eq. (3)]. That is, for an ensemble to be consistent in the context of the model defined by Eqs. (1)–(3), not only must the ensemble members exhibit μ = 0 and σ² = 1, but r = R is required also.

3. Reinterpretation of the Marzban et al. (2010) results

Marzban et al. (2010) present two experiments based on the multivariate Gaussian ensemble model described in section 2. In the first of these experiments R = r = 0.9 is specified, and RHs are shown in their slide 9 [which is also reproduced as Fig. 5 in Marzban et al. (2011)] for all combinations of the parameters μ ∈ {0.0, 0.2, 0.4, 0.8, 1.6} and σ ∈ {0.25, 0.50, 1.00, 2.00, 4.00}; analogously to Fig. 1 in Hamill (2001) who specified, in effect, R = r = 0. It is expected from the discussion in section 2 that only the RH for μ = 0.0 and σ = 1.00 should exhibit the flat signature of a consistent ensemble forecasting system, and that is indeed the case. Cases for which μ > 0 exhibit negative slopes diagnostic of overforecasting, as expected. Also expected are the U-shaped RHs for μ = 0 and σ < 1.

The unexpected result is that the two RHs for which μ = 0 and σ > 1 are also U shaped. Naively, unbiased (μ = 0) ensembles with excessive climatological variance (σ² > 1) might be expected to exhibit mound-shaped RHs characteristic of overdispersion, with the verifying observation occurring outside the ensemble too rarely. However, as shown here, these two RHs do indeed correctly diagnose ensemble underdispersion, which occurs in these cases as a result of the very strong correlation parameters R = r = 0.9, even though σ > 1.

Consider an elementary property of the multivariate Gaussian distribution [e.g., Wilks 2006b, Eq. (4.37a)], according to which the conditional mean of any ensemble member yᵢ in the multivariate model defined in section 2, given a particular value of the verifying observation, can be expressed as

$$
\mu_{y_i|y_0} = \mu_{y_i} + \frac{\sigma_{y_i}}{\sigma_{y_0}}(y_0 - \mu_{y_0}) = \mu_{y_i} + R y_0.
$$

The second equality in Eq. (4) follows from the mean and variance in Eq. (1), and the (climatological) standard deviation of the ensemble members being σ. When in addition μ = 0, the ensemble will be centered at Ry₀, which will be more extreme (farther from the origin) than y₀ for relatively large σ. That is, because R is large in slide 9 of Marzban et al. (2010), most if not all of the members of the (compact) ensemble will have the same sign as the observation y₀, but will be more extreme than y₀ because they are intrinsically more variable: the amplitudes of the excursions of the ensemble members away from the origin will be larger than that for the observation. Thus, the observation will be contained within the ensemble too rarely; the ensembles in these cases are underdispersed on average even though their climatological variance is larger than 1.

Figure 1 shows that, for the unbiased (μ = 0) forecasts in slide 9 of Marzban et al. (2010), the RMSE of the ensemble mean forecasts is substantially larger than the average sample ensemble standard deviation r̄ ens, unless σ ≈ 1, which is also diagnostic of underdispersed ensembles, in agreement with the RHs. Note that the ensembles have average standard deviation larger than 1 for σ larger than 3.162, but that these large-σ ensembles nevertheless exhibit underdispersion. Figure 1 also shows relationships between RMSE of the ensemble-mean forecast and the ensemble standard deviation for biased (μ > 0) forecasts. Not surprisingly, RMSEs for the biased forecasts are inflated relative to the μ = 0 case, even for σ ≈ 1, illustrating that this diagnostic cannot distinguish ensemble underdispersion from unconditional bias, as is also the case for the minimum spanning tree histogram (Wilks 2004). This result indicates that “inflating” ensemble variances on the basis of discrepancies with ensemble-mean MSEs (Bowler et al. 2008) should be done cautiously, accounting for any ensemble bias.

The second experiment of Marzban et al. (2010) investigates the characteristics of RHs for ensembles...
FIG. 1. Relationship of RMSE of ensemble-mean forecasts (solid curves) to the average ensemble standard deviation, for the Gaussian ensemble model in section 2 with $R = r = 0.9$. Dots on the solid curves indicate combinations of the parameters $\mu$ and $\sigma$ used in slide 9 of Marzban et al. (2010).

whose members have the same climatological distribution as the verifying observations ($\mu = 0$ and $\sigma^2 = 1$), for different combinations of $R$ and $r$. Figure 2, corresponding to the RHs in slide 10 of Marzban et al. (2010) (reproduced also as Fig. 6 in Marzban et al. 2011), shows the calibration functions for reliability diagrams characterizing probability forecasts for the event $\{y_0 \leq 0\}$. For simplicity, these probabilities have been estimated as $\Pr\{y_0 \leq 0\} \approx \sum_i I\{y_i \leq 0\}$, where $I\{\cdot\}$ is the indicator function (i.e., the “democratic voting” method), although more sophisticated probability estimators are available (e.g., Wilks 2006a). Results in each panel of Fig. 2 have been aggregated over $n = 10^6$ ensembles so that the sampling variability of the vertical positions of the points is negligible. Individual panels show results for combinations of $R$, $r \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, with the blank region in the lower-left portion of the figure corresponding to infeasible combinations of $r < R$.

As expected following the discussion in section 2, only the RHs along the main diagonal in slide 10 of Marzban et al. (2010) show flat responses, because only these cases reflect the ensemble consistency produced by $\mu = 0$, $\sigma^2 = 1$, and $r = R$. The corresponding reliability diagram calibration functions in Fig. 2 show that these ensembles also yield reliable probability forecasts, but only if the association between the ensemble members and the verifying observation, as indexed by $R$, is large enough. The deviations of these calibration functions for $R = r$ from the dashed diagonal lines are fully and quantitatively explained by sampling errors deriving from forecast probabilities being estimated from only $n_{\text{ens}} = 10$ ensemble members, which are most pronounced for the least skillful (smallest $R$) forecasts (Richardson 2001). Low-skill (small $R$) consistent ensembles would yield reliable forecast probabilities for sufficiently large $n_{\text{ens}}$ (Richardson 2001).

The panels in the upper triangle of slide 10 of Marzban et al. (2010), for which $r > R$, show U-shaped RHs that are diagnostic of underdispersed ensembles. This is a correct diagnosis because the condition $r > R$ implies that the ensemble members are more similar to each other than to the observation, on average, leading to ensembles that are too compact. This underdispersion is reflected in calibration slopes in Fig. 2 that are too shallow, which are diagnostic of overconfident probability forecasts, and fully consistent with the U-shaped rank histograms [cf. Figs. 7.8 and 7.22 in Wilks (2006b)].

There are three panels in slide 10 of Marzban et al. (2010), and in Fig. 2, for which $r < R$ and for which the parameter pair leads to a feasible covariance matrix in Eq. (3). The condition $r < R$ means that the ensemble members tend to be less similar to each other than they are to the observation, leading to overdispersed ensembles on average. The mound-shaped RHs in slide 10 of Marzban et al. (2010) correctly diagnose this overdispersion; as do the corresponding reliability diagram calibration functions in Fig. 2, which are characteristic of underconfident probability forecasts (Wilks 2006b, his Fig. 7.8).

The individual panels in Fig. 2 also report the average ensemble standard deviations and the RMSEs for ensemble-mean forecasts. The ensemble standard deviations are given by $\bar{\sigma}_{\text{ens}} = \sigma (1 - r)^{1/2} = (1 - r)^{1/2} (\text{Marzban et al. 2011})$. The ensemble standard deviations and RMSEs are consistent with the above discussion, with $\bar{\sigma}_{\text{ens}} < \text{RMSE}$ for underdispersed (overconfident) forecasts, $\bar{\sigma}_{\text{ens}} > \text{RMSE}$ for overdispersed (underconfident) forecasts, and $\bar{\sigma}_{\text{ens}} \approx \text{RMSE}$ for the consistent ensembles along the main diagonal. In these latter cases, $\bar{\sigma}_{\text{ens}}$ is slightly smaller than RMSE because the small ensemble size ($n_{\text{ens}} = 10$) produces sampling noise in the estimation of the ensemble mean, slightly inflating the RMSE. Note finally that the average ensemble standard deviation decreases monotonically with increasing $r$, independently of $R$, and is substantially smaller than the climatological ensemble standard deviation $\sigma = 1$ for large $r$.

4. Conclusions

Ensemble consistency requires not only that the univariate (“climatological”) distributions of the ensemble members and the verifying observation are the same,
but also that the joint distribution for pairs of ensemble members is the same as the joint distribution for the observation and any of the individual ensemble members. Only when this broader, multivariate, condition is met will substitution of the verifying observation for one of the ensemble members yield an unchanged multivariate joint distribution for the ensemble members.

The conclusions drawn by Marzban et al. (2010), that use of the RH should be abandoned (their slide 10), has been refuted by reinterpreting their results in light of the correct, multivariate, view of ensemble consistency. Their example RHs have been shown to correctly diagnose ensemble under and overdispersion, and bias, in the synthetic ensembles constructed from the model in section 2. Hamill (2001) and Gneiting et al. (2007) point out situations where the standard interpretations (Hamill 2001) of the RH may mislead. However, these are instances where the ensembles are not derived from a homogeneous source [Bröcker (2008) advocates homogeneity stratification of RHs], which condition does not pertain to the idealized multivariate Gaussian model of ensemble behavior presented in section 2. Similarly, random errors in the verifying observations may inflate their variability, leading to U-shaped RHs unless similar random perturbations are added to the ensemble members to simulate observation errors (Anderson 1996; Hamill 2001), thus allowing the RH to reflect ensemble consistency if it exists.

Correctly diagnosing ensemble consistency requires sensitivity to the effects of correlations among ensemble members and between the ensemble members and the observations, as well as to possible discrepancies...
between the model and observed climates. It has been demonstrated here that the RH correctly diagnoses the joint effects of both of these attributes on ensemble consistency. The assertions in Marzban et al. (2010) relating to the general unsuitability of the RH as a diagnostic tool have been shown to be false. Marzban et al. (2011) retract those initial claims regarding the RH, focusing instead on assessing the consistency of model climatologies using R-Q-Q plots. The RH continues to be a useful diagnostic for the overall consistency of ensemble forecasts.

Acknowledgments. I thank Caren Marzban for useful discussions and Tom Hamill for valuable comments on an earlier version of this manuscript.

REFERENCES


