Assessing the reliability of ensemble forecasting systems under serial dependence

Article

Accepted Version

Assessing the reliability of ensemble forecasting systems under serial dependence

Jochen Bröcker

School of Mathematical and Physical Sciences, University of Reading, United Kingdom, July 11, 2018

The problem of testing the reliability of ensemble forecasting systems is revisited. A popular tool to assess the reliability of ensemble forecasting systems (for scalar verifications) is the rank histogram; this histogram is expected to be more or less flat, since for a reliable ensemble, the ranks are uniformly distributed among their possible outcomes. Quantitative tests for flatness (e.g. Pearson's goodness–of–fit test) have been suggested; without exception though, these tests assume the ranks to be a sequence of independent random variables, which is not the case in general as can be demonstrated with simple toy examples. In this paper, tests are developed that take the temporal correlations between the ranks into account. A refined analysis exploiting the reliability property shows that the ranks still exhibit strong decay of correlations. This property is key to the analysis, and the proposed tests are valid for general ensemble forecasting systems with minimal extraneous assumptions.

Key Words: Ensemble Forecasts; Reliability; Forecast Evaluation; Rank Histograms; Serial Dependence; Statistical methods

1. Introduction

A large proportion of environmental forecasting systems nowadays issue ensemble forecasts. Such systems are used at major (national or international) weather centres, but may also form part of large scale research projects.

As with any forecasting system, there is a need to objectively assess the performance of ensemble forecasting systems. Inasmuch as ensemble forecasts provide probabilistic information about the verification, such an assessment has to be statistical in character. Several desirable (statistical) properties of ensemble (or more generally probabilistic) forecasting systems have been identified; see for instance Bröcker (2009, 2012); Weigel (2011). In the present paper, we will be concerned with *reliability*. A formal definition (in the context of ensemble forecasts) will be given in Section 3, but roughly speaking, an ensemble forecasting system is reliable if at any point n in time, the ensemble members $X_1(n), \ldots, X_K(n)$ and the verification $Y(n)$ can be considered as having been drawn independently from an underlying (or latent) forecast distribution. Reliability can be regarded as a statistical null hypothesis, and the aim of this paper is to develop tests for this null hypothesis. In essence, this means to check whether the null hypothesis is plausible given actual data, that is, an archive of verifications and corresponding ensemble forecasts.

A popular tool to assess the reliability of ensemble forecasting systems are rank histograms (see e.g. Anderson 1996; Hamill and Colucci 1997; Talagrand et al. 1997; Hamill 2001). It is assumed that the verifications are real numbers; it is therefore possible to determine, for any time instant n, the rank $R(n)$ of the verification $Y(n)$ among the ensemble members $X_1(n), \ldots, X_K(n)$. The rank $R(n)$ can assume the values $1, \ldots, K+1$, and if the ensemble forecasting system under concern is reliable, the distribution of $R(n)$ is uniform over these values. This suggests that a reliable ensemble forecasting system should produce a "more or less" uniform rank histogram. In most geophysical applications, the verification $Y(n)$ will of course not be a real number but a vector (of potentially very large dimension). There are several approaches to reduce the case of multi–dimensional verification to the scalar case (see e.g. Wilks 2004; Hansen and Smith 2004), and these can be applied without any modfication to the situation considered in the present paper. We will therefore consider the verifications to be real numbers.

In reality a rank histogram will never be precisely flat, and there are broadly speaking two possible reasons for this. Firstly, deviations from the uniform distribution might be due to the ensemble forecasting system failing to be reliable. There are certain deficiencies of ensemble forecasting systems that appear to be somewhat typical and which produce characteristic patterns in the rank histogram. A U-shaped distribution for instance indicates underdispersiveness, with a peaked distribution suggesting the opposite; sloped rank histograms show under– or overforecasting (depending on the sign of the slope).

Secondly, even a perfectly reliable ensemble forecasting system will not produce a perfectly uniform rank histogram due to random variations. Thus a test for reliability essentially amounts to a test for the hypothesis that the ranks have a discrete uniform distribution. A common test for evaluating whether a histogram is consistent with a specific discrete distribution is Pearson's goodness–of–fit (GOF) test. (Taking the ordering of the possible ranks into account, which the GOF test does not, more powerful

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/qj.3379

tests can be obtained, for instance from the Cramér–von Mises family of statistics, see Elmore (2005). In the present paper, we will focus on variants of the GOF test though.)

A serious problem with applying the GOF test directly to rank histograms for ensemble forecasting systems though is that the ranks are generally not independent. This will be demonstrated in Section 4 with a simple toy example. Independence however is an important assumption in the GOF test that can not easily be dispensed with. The general fact that verification–forecast pairs can certainly not assumed to be independent is a difficulty that affects statistical forecast evaluation in general, as has been emphasised only relatively recently (see for instance Wilks 2010; Pinson et al. 2010; Siegert et al. 2017; Bröcker 2018). A remedy suggested by Wilks (2010) is to use explicit (parametric) assumptions regarding the dependence structure and distribution of the forecasts, but the considered situation is very specific.

In the present paper, we will use an approach based on results similar to Bröcker and Kuna (2018); Bröcker (2018). The basic idea is that assuming the forecasting system is reliable, the ensemble $\mathbf{X}(n) = (X_1(n), \ldots, X_K(n))$ provides the statistical properties of $Y(n)$, given the information available at the time the forecast $\mathbf{X}(n)$ was issued, namely at time $n - L$, where L is the lead time. This fact can be used to obtain (to some extent) the statistical properties of the ranks, including their correlation structure. In fact, in certain cases (corresponding effectively to lead time $L = 1$) the ranks turn out to be independent after all, meaning that in this situation the classical GOF test can be used. In general though, the more complicated correlation structure of the ranks needs to be taken into account. We will show that this is possible, however. By modifying GOF–like tests in an appropriate manner, we obtain tests for the reliability of ensemble forecasts. These tests are valid under minimal extraneous assumptions (whereby we mean assumptions that would not automatically follow from the assumption of reliability and would have to be assumed in addition).

2. The goodness–of–fit test revisited

In this section, we will revisit the basic steps in deriving the distribution of the goodness-of-fit test statistic. In particular, we will clarify where the assumption of independence of the ranks comes in. We start with fixing some general notation. We let $\{Y(n), n = 1, \ldots, N\}$ be a series of real–valued verifications, with the index n representing the time. Further, $\{\mathbf{X}(n), n = 1, \ldots, N\}$ is a series of corresponding ensemble forecasts, where for each time instant n the ensemble is given by a vector of $K - 1$ ensemble members, that is $\mathbf{X}(n) = (X_1(n), \ldots, X_{K-1}(n))$, where each ensemble member is again real valued.[†]

For a given $y \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^{K-1}$, we consider the function $r(y, \mathbf{x})$ that is equal to k if the rank of y among the K–dimensional vector $(y, \mathbf{x})$ is equal to k. In other words, $r(y, \mathbf{x}) = k$ if precisely $k - 1$ components of $\mathbf{x}$ are smaller than or equal to y. The function r can assume the values $1, \ldots, K$. For $n = 1, \ldots, N$, we define $R(n) := r(Y(n), \mathbf{X}(n))$. That is $R(n)$ is the rank of the verification $Y(n)$ with respect to the ensemble $\mathbf{X}(n)$. We assume that the ensemble forecasting system is *reliable* with respect to the verifications. As said in the introduction, this means broadly speaking that for each time n, the verification Y_n as well as each individual ensemble member $X_k(n), k = 1, \ldots, K - 1$ can be considered independently drawn from some underlying forecast distribution. This implies (again, a proof will follow in the next section) that for each n the rank R_n is uniformly distributed over its possible values $\{1, \ldots, K\}$. As has already been mentioned though, there is no apriori reason why the ranks $R(n), n = 1, 2, \ldots$ should be independent from one another.

[†]Using $K - 1$ rather than K ensemble members will simplify subsequent notation.

To define the GOF test statistic, consider the counts

$$N_k := (\text{Number of } n \text{ for which } R(n) = k) = \sum_{n=1}^{N} \mathbb{1}_{\{R(n)=k\}},$$

where the *indicator function* $\mathbb{1}_A$ of some event A is one if the event happens and zero otherwise, and $k = 1, \ldots, K$. Clearly, the count N_k is the height of the k'th histogram bar. Further we set

$$c_k := \frac{N_k - N/K}{\sqrt{N/K}}.$$

Note that the expected value of c_k is zero, since N/K is the expected number of counts for each value of the rank, or alternatively the expected height of the k'th histogram bar. The GOF test statistic is given by

$$t = \sum_{k=1}^{K} c_k^2 = \|\mathbf{c}\|^2, \tag{1}$$

where $\mathbf{c} = (c_1, \ldots, c_K)$, and $\|.\|$ denotes the standard Euclidean norm.

As we will see now, the test statistic t has, asymptotically for large N, a χ^2 distribution with $K - 1$ degrees of freedom, if the ranks are indeed independent. The key property of the variables $c_1, \ldots, c_K$ is that they jointly satisfy a central limit theorem; for this to happen, it is sufficient that the ranks $R(n), n = 1, \ldots, N$ are independent. It is worth noting already at this point though that independence is not necessary, as will be discussed in the next section. In any event, we assume that the $c_1, \ldots, c_K$ have a joint normal distribution, with mean zero as was already noted.

We now have to calculate the covariance matrix, but before doing this, we note the following fact: let $\mathbf{v} \in \mathbb{R}^K$ be the vector with components $v_k = 1/\sqrt{K}$ for all $k = 1, \ldots, K$. Then $\|\mathbf{v}\| = 1$ and also

$$\mathbf{v}^T \mathbf{c} = \sum_{k=1}^{K} c_k v_k = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} c_k = 0.$$

If we now write $\Gamma_{i,j} := \mathbb{E}(c_i c_j)$ for the covariance matrix of $\mathbf{c}$, then

$$(\Gamma \mathbf{v})_i = \sum_{j=1}^{K} \mathbb{E}(c_i c_j) v_j = \mathbb{E}(c_i \sum_{j=1}^{K} c_j v_j) = 0.$$

This means that the nullspace (or kernel) of Γ is spanned by the constant vector $\mathbf{v}$; we stress that this is true irrespective of whether the ranks are independent or not. To find the precise shape of the covariance matrix Γ though, we have to use independence. A simple calculation will then reveal that

$$\Gamma = \mathbb{1} - \mathbf{v} \cdot \mathbf{v}^T. \tag{2}$$

This matrix is symmetric and has a nullspace spanned by $\mathbf{v}$ (as was already seen), while any other vector w with the property that $\mathbf{v}^T \mathbf{w} = 0$ is an eigenvector of Γ with eigenvalue one. The condition that $\mathbf{w}$ is perpendicular to $\mathbf{v}$ just means that $\sum_{k=1}^{K} w_k = 0$; vectors with this property are called *contrasts*.

Let now $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(K-1)}$ be a set of orthogonal contrasts (such a set can contain at most $K - 1$ elements). Then the random variables $\mathbf{d} = (d_1, \ldots, d_{K-1})$ defined through

$$d_j = \sum_{k=1}^{K} c_k w_k^{(j)} \tag{3}$$

have again a normal distribution with mean zero, but now with unit covariance matrix, since $\mathbb{E}(d_j d_k) = (\mathbf{w}^{(j)})^T \Gamma \mathbf{w}^{(k)} = \delta_{jk}$. It follows that $d_1, \ldots, d_{K-1}$ are independent standard normal.

Therefore, $\sum d_k^2$, where the index k runs over a subset of $\{1, \ldots, K-1\}$, has a χ^2 distribution, with degrees of freedom given by the size of that subset. In particular, $\|\mathbf{d}\|^2$ has a χ^2 distribution with $K-1$ degrees of freedom. But since $\|\mathbf{d}\|^2 = \|\mathbf{c}\|^2 = t$, the same is true for t.

As an aside, we note that a user has the option to assess the rank histogram by using only a subset of the random variables $d_1, \ldots, d_{K-1}$, or in other words, by projecting the scaled counts $c_1, \ldots, c_K$ onto a reduced set of contrasts. This has been suggested previously by Jolliffe and Primo (2008). The user has complete freedom in choosing the desired contrasts, as long as they are orthogonal and normalised. To obtain such a set, it is suggested to start with a set of vectors $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(\kappa)}$ that have roughly the desired shape (for instance linear, U–shaped, sinusoidal, etc) and then apply a Gram–Schmidt procedure (or equivalently a QR–decomposition, see e.g. Golub and Van Loan 1996) to the vectors $\mathbf{v}, \mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(\kappa)}$ in order to render them mutually orthogonal and normalised. Figure 1 shows three contrasts for the case of $K = 8$. These were obtained by applying a QR–decomposition to the four vectors v, $(k)_{k=1,\ldots,K}$, $(k^2)_{k=1,\ldots,K}$ and $(k^3)_{k=1,\ldots,K}$. The contrasts are linear, U–shaped and sinusoidal, respectively.

3. Tests valid under serial dependence

In the previous section, we discussed why the classical GOF test statistic has a χ^2 distribution with $K-1$ degrees of freedom. If we look back at this discussion, we find that the independence of the ranks was used in two places: in justifying a Central Limit Theorem for the $c_1, \ldots, c_K$, and when calculating the precise form of the covariance matrix Γ. With the condition of independence dropped, Γ will not have any longer the form shown in Equation (2), and this is the main reason why applying the standard GOF test to rank histograms is not warranted in general. We will discuss later in this section that a Central Limit Theorem might still hold even though the ranks are not independent. Further, even though Γ is no longer known, the relevant correlations can be estimated from the data, and an estimator will be provided below. For now, we assume that the random variables $c_1, \ldots, c_K$ have a normal distribution with mean zero and some covariance matrix Γ.

It remains true though that the nullspace of Γ is spanned by the vector $\mathbf{v}$ as the derivation of this fact in the previous section did not depend on independence of the ranks. This implies that we still get a faithful representation of the scaled counts $c_1, \ldots, c_K$ by projecting then onto a set of orthonormal contrasts as in Equation (3), that is by using the random variables $d_1, \ldots, d_{K-1}$ defined through Equation (3). We want to develop a test based on a subset $\mathbf{d} = (d_1, \ldots, d_\kappa)$ of these random variables, and we denote the covariance matrix of these random variables by $\Upsilon_{i,j} = \mathbb{E}(d_i d_j) = (\mathbf{w}^{(i)})^T \Gamma \mathbf{w}^{(j)}$, where $i, j \leq \kappa \leq K-1$. We keep κ fixed throughout the remainder of this section.

As the condition of independence of the ranks has been dropped, Υ will not be the unit matrix any longer. (We note again that Υ will later have to be estimated from the data.) We consider the statistic $t_\kappa = \mathbf{d}^T \Upsilon^{-1} \mathbf{d}$. This statistic is indeed a generalisation of the statistic t from the previous section, and the two agree if the ranks are independent and $\kappa = K-1$. Our claim is that t_κ has a χ^2–distribution with κ degrees of freedom as in the independent case.

To see this, let U be a symmetric matrix so that $U\Upsilon U = \mathbb{1}$ (i.e. U is a square root of Υ^{-1}). Then $\mathbf{e} = U\mathbf{d}$ is a vector of normal random variables with zero mean and covariance matrix $U\Upsilon U = \mathbb{1}$, hence the components of $\mathbf{e}$ are independent and standard normal. As a consequence, $\tilde{t} = \|\mathbf{e}\|^2$ has a χ^2–distribution with κ degrees of freedom. However,

$$\tilde{t} = \|\mathbf{e}\|^2 = \mathbf{e}^T \mathbf{e} = \mathbf{d}^T U \cdot U \mathbf{d} = \mathbf{d}^T \Upsilon^{-1} \mathbf{d} = t_\kappa,$$

proving our claim.

For the remainder of this section, we will fill in the missing parts of our argument. We will show that although the ranks are not independent, they nevertheless satisfy a very strong decay of correlation property which is a direct consequence of the reliability assumption and forms the core of our analysis. We then provide an estimator of the covariance matrix Υ. The feasibility of this estimator is due to the strong decorrelation property of the ranks, and the assumption that the ranks form a stationary and ergodic sequence. Stationarity of the sequence $(R(1), R(2), \ldots)$ means that for any m, the joint distribution of $(R(n), \ldots, R(n+m))$ does not depend on n or, roughly speaking, is invariant with respect to temporal shifts. A stationary sequence is ergodic if any average of the form

$$\frac{1}{N} \sum_{n=1}^{N} \phi(R(n), \ldots, R(n+m)) \qquad (m \text{ fixed})$$

converges to $\mathbb{E}[\phi(R(n), \ldots, R(n+m))]$ as $N \to \infty$. Note that by stationarity, this quantity does not depend on n. As ergodicity usually presumes stationarity, we will take "ergodic" to mean "stationary and ergodic". Ergodicity of the ranks is the only extraneous assumption we need to add. These properties are also sufficient to justify the validity of the Central Limit Theorem (more details will be provided in Appendices A and B).

The reliability assumption is interpreted to mean the following. For every time instant $n = 1, \ldots, N$ there exists an underlying or latent forecast distribution μ_n over the real numbers. This distribution is itself random and represents the distribution of Y_n conditional on the information available at forecast time. More formally, let $\mathcal{F}_n$ be the information available to the forecaster at time n, and say that forecasts are issued with a lead time L, then reliability means that

$$\mu_n(A) = \mathbb{P}(Y(n) \in A | \mathcal{F}_{n-L})$$

for all $n = 1, \ldots, N$ and any set A on the real line.[‡] The joint set of verification and ensemble members $(Y_n, X_1(n), \ldots, X_{K-1}(n))$ are independently drawn from this distribution, that is, for any n and any sets $A_0, \ldots, A_{K-1}$ on the real line, it holds that

$$\mathbb{P}(Y(n) \in A_0, X_1(n) \in A_1, \ldots, X_{K-1}(n) \in A_{K-1} | \mathcal{F}_{n-L}) = \mu_n(A_0) \cdot \ldots \cdot \mu_n(A_{K-1}).$$

The uniform distribution of the ranks, *conditional* on the forecast information, is now an elementary consequence: for all $n = 1, \ldots, N$ and $k = 1, \ldots, K$ we have

$$\mathbb{P}(R(n) = k | \mathcal{F}_{n-L}) = \frac{1}{K}. \tag{4}$$

We will graft another element to the reliability assumption which is usually not made explicit but is evidently satisfied in most applications, namely that for any n, the forecast information $\mathcal{F}_n$ contains all verifications and ensembles up to that point; in other words, at any time n the forecaster knows $\{Y(m), m = 1, \ldots, n\}$ and also $\{\mathbf{X}(m), m = 1, \ldots, n\}$. This, in combination with Equation (4), yields the following key identity:

$$\mathbb{P}(R(n) = k | R(1), \ldots, R(n-L)) = \frac{1}{K} \tag{5}$$

[‡] Strictly speaking for any measurable set A on the real line.

for all $n = 1, \ldots, N$ and $k = 1, \ldots, K$. Another way of saying this is that for any n, the rank R_n is uniformly distributed and independent from the ranks $R(1), \ldots, R(n-L)$, that is, from the ranks known at forecast time. In particular, we obtain that in the case of unit lead time (i.e. $L = 1$), the ranks $\{R(n), n = 1, 2, \ldots\}$ are indeed fully independent; this implies that in this special (but important) situtation, the classical GOF test for the rank histogramm is valid.

Let now $\{\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(\kappa)}\}$ be a set of orthonormal contrasts, and define

$$Z_k(n) = \sqrt{K} \sum_{j=1}^{K} w_j^{(k)} \mathbb{1}_{\{R(n)=j\}} \tag{6}$$

for $n = 1, \ldots, N$ and $k = 1, \ldots, \kappa$; note that $d_k = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} Z_k(n)$. We regard $\mathbf{Z}(n) = (Z_1(n), \ldots, Z_\kappa(n))$ with $n = 1, \ldots, N$ as a sequence of random vectors. The property (5) implies that this sequence has finite correlation length of at most $L - 1$. To see this, note that for any n, the random vector $\mathbf{Z}(n)$ depends on $R(n)$ only. Hence, $\mathbf{Z}(n+l)$ is independent of $\mathbf{Z}(n)$ if $l \geq L$. Further, $\mathbb{E}(\mathbf{Z}(n)) = 0$ and therefore

$$\mathbb{E}(\mathbf{Z}(n+l) \cdot \mathbf{Z}(n)^T) = \mathbb{E}(\mathbf{Z}(n+l)) \cdot \mathbb{E}(\mathbf{Z}(n)^T) = 0. \tag{7}$$

It turns out that in order to establish a joint Central Limit Theorem for $\mathbf{d} = (d_1, \ldots, d_\kappa)$, an additional assumption is needed, namely that the ranks $\{R(n), n = 1, 2, \ldots\}$ form a stationary and ergodic sequence. With this assumption and property (7) in place, it follows from established results that $\mathbf{d}$ will be asymptotically normal with mean zero and some covariance matrix Υ; we will not provide a proof here, but some more details and references can be found in Appendix B.

An estimator for Υ, the asymptotic covariance matrix of $\mathbf{d}$, is needed as well. We will use the estimator

$$\Upsilon_N = \mathbb{1} + \frac{1}{N} \sum_{n=1}^{N} \sum_{l=1}^{L-1} \left\{ \mathbf{Z}(n)\mathbf{Z}(n+l)^T + \mathbf{Z}(n+l)\mathbf{Z}(n)^T \right\}. \tag{8}$$

This estimator can be shown to converge to Υ, and a demonstration can be found in Appendix A. We stress that the validity of this estimator rests not only on the ergodicity assumption but also on the finite correlation property (7). For the case $L = 1$, this estimator reduces to $\Upsilon_N = \mathbb{1}$ as it should.

4. Numerical examples

We start this section with a short list summarising the steps needed to perform the test for flatness of a rank histogram. We let $\{(Y(n), \mathbf{X}(n)), n = 1, \ldots, N\}$ be a sequence of real–valued verifications and corresponding ensembles with $K - 1$ members. Let further $\{\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(\kappa)}\}$ be a set of orthonormal contrasts, describing possible deviations of a rank histogram from flatness (with $\kappa \leq K - 1$).

1. Compute the ranks $\{R(n), n = 1, \ldots, N\}$.
2. Using the ranks and the contrasts, compute $Z_k(n)$ from Equation (6) for $n = 1, \ldots, N$ and $k = 1, \ldots, \kappa$.
3. Compute the estimator Υ_N for the covariance Υ from Equation (8).
4. Compute $d_k = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} Z_k(n)$ for $k = 1, \ldots, \kappa$ and let $\mathbf{d} = (d_1, \ldots, d_\kappa)$.
5. Now $\mathbf{d}^T \Upsilon_N^{-1} \mathbf{d}$ should have a χ^2 distribution with κ degrees of freedom, and this can be used to compute the p–value.

For the remainder of this section, we will discuss two numerical examples. The first example considers a simple autoregressive process; this has been chosen merely to illustrate the methodology. The second example uses data from an assimilation experiment using the two dimensional Navier–Stokes equation.

Example 1: Autoregressive process In the first example, the verification $\{Y_n, n = 1, 2, \ldots\}$ forms an autoregressive (AR) process of the form

$$Y(n+1) = \alpha Y(n) + \zeta(n+1), \tag{9}$$

where $\{\zeta(n), n \in \mathbb{Z}\}$ is a sequence of independent standard normal random variables and $\alpha = 0.95$. The information $\mathcal{F}_n$ available to the forecaster at time n is $\{Y(k), k \leq n\}$, that is the entire history of observations up to and including $Y(n)$. Reliable ensemble forecasts can be generated by replacing $\zeta(n)$ in Equation (9) with independent realisations of the noise process. More specifically, let $\{\boldsymbol{\xi}(n), n = 1, 2, \ldots\}$ be a sequence of independent random vectors $\boldsymbol{\xi}(n) = (\xi_1(n), \ldots, \xi_{K-1}(n))$, where the components $\xi_k(n)$ are again independent and standard normal. Then an ensemble forecast for lead time L and verifying at time $n + L$ is given by

$$X^k(n+L) = \alpha^L Y(n) + \sigma_L \xi_k(n), \qquad k = 1, \ldots, K-1;$$

here, $\sigma_L^2 = \frac{1-\alpha^{2L}}{1-\alpha^2}$.

In this model, it is easy to see directly that two ranks are independent if they are L or more steps apart, but that they are dependent otherwise. To check this, we write $Y(n+L)$ as

$$Y(n+L) = \alpha^L Y(n) + \sum_{l=0}^{L-1} \alpha^l \zeta(n+L-l). \tag{10}$$

Therefore,

$$\begin{aligned} R(n+L) &= r(Y(n+L), \mathbf{X}(n+L)) \\ &= r(\sum_{l=0}^{L-1} \alpha^l \zeta(n+L-l), \sigma_L \boldsymbol{\xi}(n)). \end{aligned} \tag{11}$$

(We recall that $r(y, \mathbf{x})$ is the rank of y among the components of $\mathbf{x}$.) Equation (11) demonstrates that the temporal dependence of the ranks is due to the temporal dependence of $\iota_L(n) := \sum_{l=0}^{L-1} \alpha^l \zeta(n+L-l)$. In view of Equation (10), the random variable $\iota_L(n)$ describes the subsequent evolution of the observations after the forecast $\mathbf{Z}(n)$ has been issued. We might call $\iota_L(n)$ the *innovation*; it is precisely the part of $Y(n+L)$ not captured by the forecast. If two observations $Y(n)$ and $Y(m)$ are less than L time steps apart (i.e. $|m-n| < L$), then their corresponding innovations will be dependent, due to overlap of their evolutions after the respective forecasts have been issued. This is also evident from the expression of the innovation. If $|m-n| \geq L$ though, their innovations will be independent. Due to Equation (11), the ranks will exhibit the same phenomenon.

Figure 2 shows typical histograms for ensemble forecasts in the context of the AR process. The ensemble forecasting system uses 7 members, and the data set comprised 100 time instances. The lead time was 1 time unit for the top panel and 10 time units for the bottom panel of Figure 2. It is evident that the histogram for the larger lead time shows considerably stronger variations in the counts. This is due to the strong temporal correlations between the ranks at larger lead times. The p–values for the top and bottom panels are 0.7612 and 0.7199, respectively, using the test proposed in Section 3 for the second histogram. Using a classical GOF test would give a p–value of 0.0019 for the second histogram, thus concluding wrongly that this forecast is not reliable.

In order to check whether the test presented in Section 3 takes the correlations correctly into account, we have created 1,000 Monte Carlo resamples of the experiment described above, albeit with 400 time instances. For every Monte Carlo sample, we computed the statistic t_κ for $\kappa = 2$, using a linear and a U–shaped contrast, as described in Section 3, including the estimator

of the covariance matrix. If the presented theory is correct, then t_κ should follow a χ^2 distribution with 2 degrees of freedom, or equivalently the p–value should have a uniform distribution. This turns out to be the case; a histogram of the p–values obtained from our 1,000 Monte Carlo resamples is shown in the top panel of Figure 3. Furthermore, a Kolmogorov–Smirnov test yields a p–value of 0.6876, confirming that these follow a uniform distribution.

For each Monte Carlo resample we have also calculated the *classical* GOF statistic, that is, ignoring the correlations in the ranks and assuming that Υ is the identity matrix. That the resamples of that statistic do *not* follow a χ^2 distribution with 2 degrees of freedom is evident from the bottom panel of Figure 3, which shows a histogram of the p–values. These are evidently concentrated at too low values, which implies that ignoring the correlations in the ranks and applying the classical GOF test would result in too frequent rejection, that is, we would conclude too often that the rank histogram is not consistent with reliability.

It should be said that the bottom panel of Figure 3, although not inconsistent with reliability, is a somewhat extreme case. Using a single (U–shaped) contrast, we obtain a test that is more powerful against U–shaped deviations from reliability. Applying this test, the p–value of the example in Figure 3, bottom panel, becomes 0.0532 and is thus on the verge of being rejected with this test.

Regarding the power of the full test (using all contrasts), we might wonder how large N, the number of time instances, would have to be in order that a p–value of 0.05 or less is obtained, while the observed relative frequencies as well as the covariance matrix Υ_N being left the same. It is easy to see that in this situation, the test statistic $\tilde{t}$ is proportional to N; using the current value of $\tilde{t} = 4.5072$ and the inverse cumulative distribution function of the χ^2 distribution with $K-1$ degrees of freedom, we find that N has to increase about threefold (i.e. to about 300) to reject the histogram in Figure 3, bottom panel, as not flat.

We have also investigated the role of α and its influence on the rank correlations. Strictly speaking, we should investigate the correlation structure of the $\mathbf{Z}(n), n = 1, 2, \ldots$, as this will determine the magnitude of Υ, and the larger this quantity the less powerful the test will be. To simplify the discussion though, we focus on the correlation of the ranks directly; in the case of a single linear contrast, these are in fact sufficient to determine Υ. Figure 4 shows the Pearson correlation coefficient

$$\rho_m = \frac{\mathrm{Cov}(R_n, R_{n+m})}{\mathrm{Var}(R_n)}.$$

between the ranks R_n and R_{n+m} for ensemble forecasts for the AR process, for values of $m = 0, \ldots, 10$ on the abscissa and several values of α (marked with different graphic symbols, see figure). A simple calculation, not shown here, reduces calculation of correlation coefficient to the numerical evaluation of an integral. (It turns out that the correlation coefficient does not depend on the number of ensemble members.) As discussed, the rank correlation is zero for $m \geq L$ irrespective of α. Nonetheless, the correlation for $m < L$ depends on α and decreases faster for smaller values of α. It is also easy to see directly that the ranks are once again independent in the limiting case $\alpha = 0$.

The (probably not surprising) conclusion is that the dependence structure of the ranks depends both on the lead time L as well as the dependence structure of the verification–forecast pairs which is ultimately determined by the nature of the uncerlying problem. In particular, while the lead time L provides an upper bound on the maximal correlation length of the ranks, fast decorrelation of the verification–forecast pairs can render the correlation for larger lags very small or even negligible. Furthermore, fast decorrelation of the verification–forecast pairs will, in general, lead to the test being more powerful. Prior knowledge about the correlation structure of the verification–forecast pairs might be used to further increase the efficiency of the estimator for Υ, but it is not clear how to do that in an operational situation and whether the additional efford required would pay off in terms of increased test power.

Example 2: Data assimilation in 2D Navier–Stokes The second example uses data from an assimilation experiment with the two dimensional Navier–Stokes equation. The equation was implemented in the vorticity–streamfunction formulation

$$\partial_t \omega + J(\omega, \psi) + A\omega = f, \tag{12}$$

on the two–dimensional unit torus $\mathbb{T} =]0, 1[^2$ with periodic boundary conditions. Here, ω is the vorticity and ψ the stream function; further, $A = -\nu\Delta$ (the Laplacian with viscosity ν), and the stream function is obtained from the vorticity through solving the Poisson equation $\Delta\psi = \omega$. The function f represents a forcing. Equation (12) (along with the Poisson equation) was solved with a pseudospectral code on a square spatial lattice with resolution $N = 21$ in both dimensions. In other words, the equation was truncated at wavenumber 10, where we define the *wavenumber* of a wave vector (k, l) as $|(k, l)| := \max\{|k|, |l|\}$. The viscosity was set to $\nu = 2 \cdot 10^{-3}$. The forcing was time independent and composed of randomly selected amplitudes and truncated at wavenumber 3, with a magnitude of $\|f\| = 1.34$. In this setup, the system produces complex nonperiodic solutions. (Here and in the following, we use the norm $\|f\| = \left(\int_{\mathbb{T}} |f|^2(x)\mathrm{d}x\right)^{1/2}$ for a—possibly complex—function on the torus.)

Observational data was assimilated into an identical copy of the two dimensional Navier–Stokes equation. As observations, the Fourier modes with wavenumbers $|(k, l)| \leq 1$ were used (which corresponds to observing nine modes, or equivalently, to taking smoothed spatial observations on a grid with 3×3 gridpoints). The observations were taken at temporal intervals of $\Delta t = 0.5$ time units and corrupted with normally distributed noise of about 5%. The observations were then assimilated simply by replacing the relevant Fourier modes of the assimilated solutions with the observed Fourier modes (see Hayden et al. 2011; Sanz Alonso and Stuart 2014; Bröcker et al. 2017, for theoretical analyses of this assimilation method).

Ensembles were generated by randomly perturbing the analyses fields. The distribution of the perturbations was taken to be normal with mean zero and standard deviation $\|\delta\omega\| = 0.943$. Ensembles were generated by integrating the model forward with these perturbed analysis fields serving as initial conditions. The standard deviation for the perturbations was found by optimising the mean square forecast performance for lead time of 5 units in an offline experiment.

As verifications in these experiments, we use one of the nine components of the observations employed for the data assimilation (recall that observations on a 3×3 grid were used for data assimilation). We analysed these verifications and corresponding ensembles for lead times of $L = 5$, 10 and 20 time units, each data set comprising 300 verification–forecast pairs. No attempt was made to statistically recalibrate these ensembles. Although there is no model error in this experiment, this does not imply that the ensemble forecasting system is reliable, since the data assimilation system is fairly primitive and we have no reason to believe that ensembles comprise a reasonable representation of the forecast distribution. The histograms for these three data sets are shown in Figure 5. It is seen that the reliability of this relatively simple ensemble forecasting system is not bad by visual inspection. We applied the described test for flatness

of the rank histogram, first for two contrasts (i.e. $\kappa = 2$). The p–values for lead times $L = 5$, 10 and 20 are 0.7872, 0.7495, and 0.5209, respectively. Testing the full set of contrasts gives p–values of 0.5507, 0.5572, and 0.5854; all these figures do not provide a strong case for deviation from reliability. With regards to the rank histogram corresponding to lead time $L = 20$ in particular though, the histogram appears to have a slight slant to the right (indicating underforecasting), but this effect might be masked by the expected variation of the histogram. However, we find that $\mathrm{trace}(\Upsilon_N) = 8.63$, while this value would be 7 for independent ranks, and we can conclude that the variance of the histogram is not in fact much larger than for the independent case.

We repeated the test for lead time $L = 20$ with a single, linear contrast and find a p–value of 0.3254, which might indicate a slight deviation from reliability. Note that we have cheated a little bit, as the choice of the contrast was made based on the data; choosing the contrasts depending on the data means that the contrasts would be functions of the data while the testing methodology assumes they are not. As a final note, under the assumption of uncorrelated ranks the p–value for this case would have been 0.2676, so not in fact very different. For the variance, we have the estimate 1.2708 which is fairly close to 1, again indicating that dropping the assumption of independence does not make much of a difference in this case.

From our discussion of the AR–process, we speculate that this is due to a relatively fast decay of temporal dependencies in the verification–forecast pairs, which would imply that although the correlations in the ranks cannot extend beyond lag 20 in this case, they are effectively much shorter in the present situation. Figure 6 shows the estimated correlation coefficient between the ranks R_n and R_{n+m} for this ensemble forecasting system at a lead time of 20 time units; values of the lag m between zero and 20 are shown on the abscissa. This is just an estimate of the correlation and although we have ommitted any uncertainty information such as error bars, there is no question that the correlation decreases indeed very quickly with increasing lag, and correlation with larger lag do not contribute much to Υ, due to fast decay of correlation in this system. This implies that the properties of the test in this example are very similar to the standard GOF test. This is true in general if the estimator Υ_N for the covariance matrix is observed to be close to the unit matrix (as is the case in the present example), which is easy to check in applications.

5. Conclusions and outlook

A popular and practical tool to assess the reliability of ensemble forecasting systems (for scalar verifications) is the rank histogram. For a reliable ensemble forecasting system, this histogram is expected to be more or less flat, since the ranks are uniformly distributed among their possible outcomes. For a more quantitative analysis though, it would be desirable to have a test for the flatness of rank histograms, as certain random fluctuations will always be present even if the forecasting system is reliable. We have argued that classical approaches such as for example Pearson's goodness–of–fit test are not appropriate since these tests rest on the assumption that the ranks form a sequence of independent random variables. By revising the derivation of Pearson's goodness–of–fit test, we identified two places where the assumption of independence is relevant: firstly it ensures that the rescaled histogram counts satisfy a joint Central Limit Theorem, and secondly it entails a very specific correlation structure for these counts.

Although the ranks of a reliable ensemble forecasting system are not independent in general, we have demonstrated both analytically and numerically that an appropriate modification of the goodness–of–fit test will still work. Central to our analysis is the fact that for a reliable ensemble forecasting system, the ranks still satisfy a strong decay of correlation property—the correlation time of the ranks is even finite and given by the lead time less one. (This result can be generalised to different types of forecasting systems and might be of independent interest, see Bröcker and Kuna (2018); Bröcker (2018).) Furthermore, it was shown how to perform a "reduced" goodness–of–fit test using a restricted set of contrasts, as suggested in Jolliffe and Primo (2008), but modified so as to account for rank correlations. Apart from the technical condition that the ranks form an ergodic sequence, the approach does not require any extraneous or distributional assumptions.

The formalism was also applied to numerical examples. First, data from a simple autoregressive process was considered, with ensemble forecasts that were by construction reliable. The experiments confirm that the formalism gives the correct results, while not taking the rank correlations into account (by using a classical goodness–of–fit test) yields too high rejection rates as the distribution of the classical goodness–of–fit test statistic is no longer a χ^2–distribution.

A second example used data from a simple fluid dynamical data assimilation experiment. The results show that despite a relatively crude data assimilation system, the ensembles are fairly reliable. We also addressed the question whether the test looses power for longer lead times as potentially systematic deviations from a flat rank histogram are masked by strong variability of the histogram counts, which seems not the case in that situation.

Outlook and future work An important fact emerging from our analysis is that for a reliable ensemble forecasting system, the ranks exhibit a finite correlation time which cannot exceed the lead time. This result can be generalised to different types of forecasting systems as has been done in Bröcker and Kuna (2018); Bröcker (2018). Strong decay of correlations though typically implies powerful asymptotic limit results such as Laws of Large Numbers and Central Limit Theorems. It seems plausible that these can be exploited to analyse other forecast evaluation techniques rigorously under serial dependence; examples are reliability diagrams (Bröcker and Smith 2007) or Receiver (or Relative) Operating Characteristic (Egan 1975; Bröcker 2012).

An extension of the results in the present paper to *stratified* rank histograms would also be desirable (Siegert et al. 2012). Stratified rank histograms provide a more detailed picture of reliability, conditional on different forecasting situations. This extension seems to be fairly immediate and will be dealt with in a forthcoming paper.

A. Covariance estimator

In this appendix, we discuss an estimator for Υ, the covariance matrix of $\mathbf{d} = (d_1, \ldots, d_{K-1}) = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbf{Z}(n)$ in the limit $N \to \infty$, that is

$$\Upsilon = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[\left(\sum_{n=1}^{N} \mathbf{Z}(n) \right) \left(\sum_{n=1}^{N} \mathbf{Z}(n) \right)^T \right].$$

(Notation and definitions are as in Sec. 3.) We start with studying the (matrix valued) covariance function

$$\gamma(l) := \mathbb{E}(\mathbf{Z}(n)\mathbf{Z}(n+l)^T),$$

noting that there is no dependence on n since $\{\mathbf{Z}(n), n = 1, 2, \ldots\}$ is assumed ergodic and thus in particular stationary; note also that $\gamma(l)$ is defined for negative l, too, and in fact $\gamma(-l) = \gamma(l)^T$. Furthermore, we have $\gamma(l) = 0$ if $l \geq L$ due to Equation (7). An

elementary calculation then gives

$$\frac{1}{N}\mathbb{E}\left[\left(\sum_{n=1}^{N}\mathbf{Z}(n)\right)\left(\sum_{n=1}^{N}\mathbf{Z}(n)\right)^{T}\right] = \sum_{l=-N+1}^{N-1}(1-\frac{|l|}{N})\gamma(l)$$

and hence

$$\Upsilon = \lim_{N\to\infty}\sum_{l=-N+1}^{N-1}(1-\frac{|l|}{N})\gamma(l) = \sum_{l\in\mathbb{Z}}\gamma(l). \tag{13}$$

Thanks to Equation (7), the sum in Equation (13) contains only finitely many nonzero terms, namely for $|l| < L$. These terms can be estimated by empirical averages (i.e. averages over time), that is

$$\gamma_N(l) = \frac{1}{N}\sum_{n=1}^{N}\mathbf{Z}(n)\mathbf{Z}(n+l)^{T},$$

which converges to $\gamma(l)$ due to the condition that the ranks are ergodic (we only need estimators for $0 < l < L$ since $\gamma(-l) = \gamma(l)^{T}$ is symmetric and $\gamma(0)$ is the unit matrix). The estimator Υ_N for Υ is given by replacing $\gamma(l)$ in Equation (13) with the estimators $\gamma_N(l)$. This gives

$$\begin{aligned}\Upsilon_N &= \mathbb{1} + \sum_{l=1}^{L-1}\gamma_N(l) + \gamma_N(l)^{T} \\ &= \mathbb{1} + \frac{1}{N}\sum_{n=1}^{N}\sum_{l=1}^{L-1}\mathbf{Z}(n)\mathbf{Z}(n+l)^{T} + \mathbf{Z}(n+l)\mathbf{Z}(n)^{T}.\end{aligned} \tag{14}$$

B. The Central Limit Theorem

In this appendix, we justify the a joint Central Limit Theorem for $d = (d_1, \ldots, d_{K-1})$, where $d_k = \frac{1}{\sqrt{N}}\sum_{n=1}^{N} Z_k(n)$. By a classical argument known as the Cramér–Wold device in probability theory (see for instance van der Vaart 2000, pg.16) it is sufficient to establish a central limit theorem for $\delta := \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\Lambda(n)$ where $\Lambda(n) := \boldsymbol{\lambda}^{T}\mathbf{Z}(n)$ for any vector $\boldsymbol{\lambda} \in \mathbb{R}^{K-1}$, thereby reducing the problem from a vector valued to a single valued Central Limit Theorem. Our assumptions and the discussion in the previous appendix entail that $\{\mathbf{Z}(n), n = 1, 2, \ldots\}$ are ergodic and have summable correlations. The same is therefore true for $\{\Lambda(n), n = 1, 2, \ldots\}$, and we can apply Theorem 4.18 in van der Vaart (2010) to conclude that the distribution of δ is asymptotically normal. In summary, we obtain the required joint Central Limit Theorem for $(d_1, \ldots, d_{K-1})$.

References

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

Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643): 1512 – 1519, 2009.

Jochen Bröcker. Probability forecasts. In Jolliffe and Stephenson (2012), chapter 8, pages 119–139.

Jochen Bröcker. Towards a framework for the statistical evaluation of forecasting systems under serial dependence. Technical report, Department of Mathematics and Statistics, University of Reading, 2018.

Jochen Bröcker and Tobias Kuna. On the statistical evaluation of forecasting systems. Technical report, Department of Mathematics and Statistics, University of Reading, 2018.

Jochen Bröcker and Leonard A. Smith. Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 22(3):651–661, June 2007.

Jochen Bröcker, Tobias Kuna, and Lea Oljača. Almost sure error bounds for data assimilation in dissipative systems with unbounded observation noise. 2017. (submitted).

James P. Egan. *Signal detection theory and ROC analysis*. Academic Press series in cognition and perception. Academic Press, first edition, 1975.

Kimberly L. Elmore. Alternatives to the chi-square test for evaluating rank histograms from ensemble forecasts. *Weather and Forecasting*, 20(5):789–795, October 2005.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. The Johns Hopkins University Press, 3rd edition, 1996.

Thomas M. Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3):550–560, 2001.

Thomas M. Hamill and Stephen J. Colucci. Verification of Eta–RSM short range ensemble forecasts. *Monthly Weather Review*, 125:1312–1327, 1997.

J.A. Hansen and L.A. Smith. Extending the limits of forecast verification with the minimum spanning tree. *Monthly Weather Review*, 132(6):1522–1528, 2004.

Kevin Hayden, Eric Olson, and Edriss S. Titi. Discrete data assimilation in the lorenz and 2d navier–stokes equations. *Physica D: Nonlinear Phenomena*, 240(18):1416 – 1425, 2011. doi: http://dx.doi.org/10.1016/j.physd.2011.04.021.

Ian T. Jolliffe and Cristina Primo. Evaluating rank histograms using decompositions of the chi-square test statistic. *Monthly Weather Review*, 136(6):2133–2139, 2008. doi: 10.1175/2007MWR2219.1.

Ian T. Jolliffe and David B. Stephenson, editors. *Forecast Verification; A practicioner's Guide in Athmospheric Science*. John Wiley & Sons, Ltd., Chichester, second edition, 2012.

Pierre Pinson, Patrick McSharry, and Henrik Madsen. Reliability diagrams for non2010parametric density forecasts of continuous variables: Accounting for serial correlation. *Quarterly Journal of the Royal Meteorological Society*, 136(646):77–90, 2010. doi: 10.1002/qj.559. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.559.

Daniel Sanz Alonso and Andrew M. Stuart. Long-time Asymptotics of the Filtering Distribution for Partially Observed Chaotic Dynamical Systems. *ArXiv e-prints*, November 2014.

Stefan Siegert, Jochen Bröcker, and Holger Kantz. Rank histograms of stratified monte-carlo ensembles. *Quarterly Journal of the Royal Meteorological Society*, 140(12):1558–1571, 2012. doi: http://dx.doi.org/10.1175/MWR-D-11-00302.1.

Stefan Siegert, Omar Bellprat, Martin Ménégoz, David B. Stephenson, and Francisco J. Doblas-Reyes. Detecting improvements in forecast correlation skill: Statistical testing and power analysis. *Monthly Weather Review*, 145 (2):437–450, 2017.

Olivier Talagrand, R. Vautard, and B. Strauss. Evaluation of probabilistic prediction systems. In *Workshop on Predictability*, pages 1–25. ECMWF, 1997.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

Aad W. van der Vaart. Time series, 2010. lecture notes.

Andreas P. Weigel. Verification of ensemble forecasts. In Jolliffe and Stephenson (2012), chapter 9, pages 141–166.

Daniel S. Wilks. The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. *Monthly Weather Review*, 132(6): 1329–1340, 2004.

Daniel S. Wilks. Sampling distributions of the Brier score and Brier skill score under serial dependence. *Quarterly Journal of the Royal Meteorological Society*, 136(653):2109–2118, 2010. ISSN 1477-870X. doi: 10.1002/qj.709. URL `http://dx.doi.org/10.1002/qj.709`.

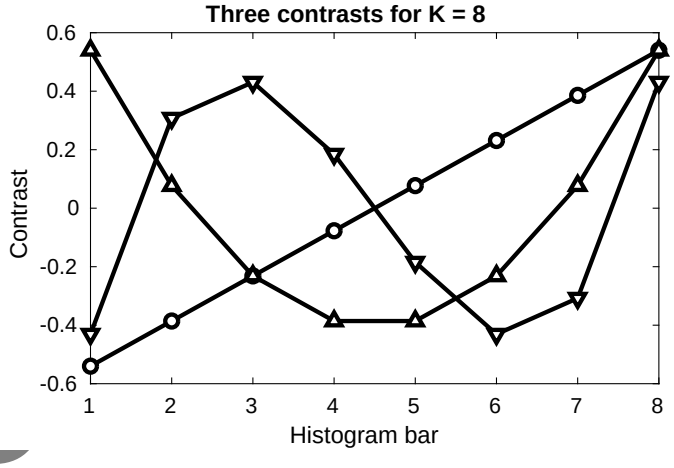

Figure 1. The figure shows three contrasts for the case of $K = 8$. These were obtained by applying a QR–decomposition to the four vectors v, $(k)_{k=1,\ldots,K}$, $(k^2)_{k=1,\ldots,K}$ and $(k^3)_{k=1,\ldots,K}$. The contrasts are linear, U–shaped and sinusoidal, respectively. (Lines connecting the points are merely for guidance.)

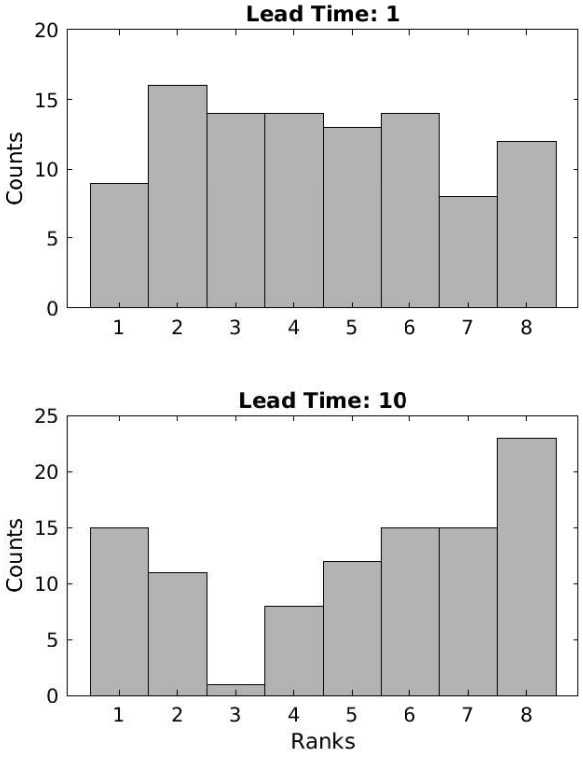

Figure 2. Typical histograms for ensemble forecasts for the AR process. The ensemble had 7 members, and the data set comprised 100 time instances. The lead time was 1 time unit for the top panel and 10 time units for the bottom panel. Althoug both forecast systems are by construction reliable, the histogram for the larger lead time is considerably "rougher", that is there are stronger variations in the counts. This is due to the positive temporal correlations between the ranks for the forecasting system at larger lead times.

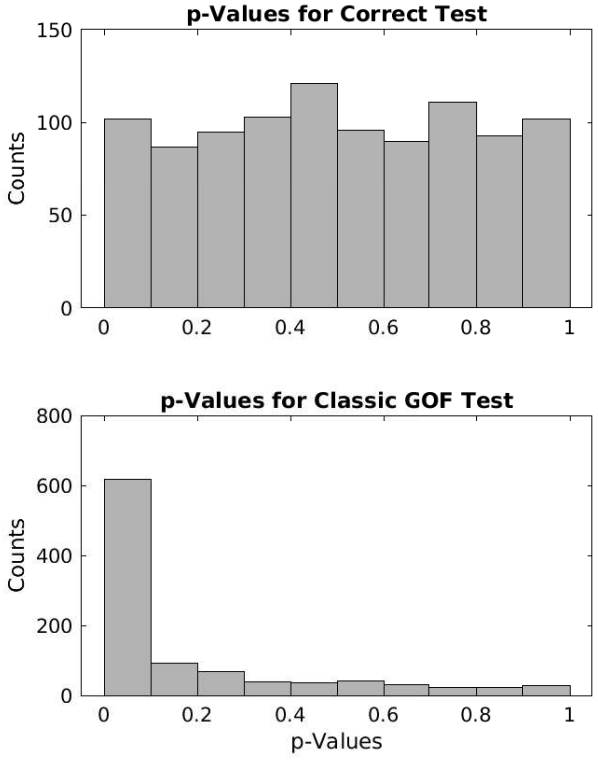

Figure 3. Histograms of the p–values of testing flatness of the rank histograms for the AR process at lead time 10. The ensemble had 7 members, and the data set comprised 400 time instances. The test statistic employed two contrasts (linear and U–shaped). The p–values were obtained from 1000 Monte Carlo repetitions of the same experiment. The top panel shows the p–values from the new test proposed in Section 3 taking the rank correlations into account. The bottom panel shows the p–values from a classical GOF test. It can be seen that the new test produces correct p–values, while ignoring the rank correlation results in too low p–values and thus too frequent rejection of the null hypothesis.

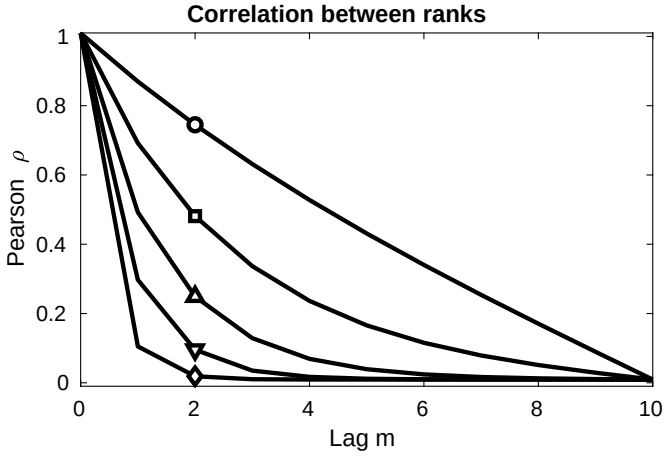

Figure 4. Pearson correlation coefficient between the ranks R_n and R_{n+m} for ensemble forecasts for the AR process. The lead time in this case was ten time units; values of the lag m between zero and ten are shown on the abscissa. Different values of α are marked with different graphic symbols: 0.1 ($\Diamond$), 0.3 ($\bigtriangledown$), 0.5 ($\bigtriangleup$), 0.7 ($\square$), 0.9 ($\bigcirc$). The correlation coefficient does not depend on the number of ensemble members. As discussed, the rank correlation is zero for $m \geq L$ irrespective of α. Nonetheless, the correlation for $m < L$ depends on α and decreases faster for smaller values of α.

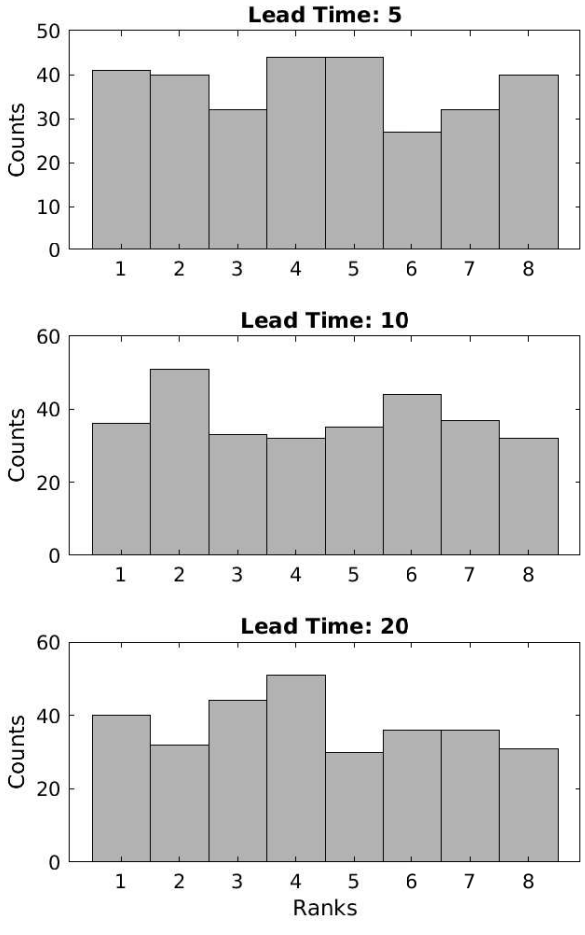

Figure 5. Rank histograms for lead times of $L = 5$, 10 and 20 in the Navier–Stokes experiment (top, middle, and bottom panel, respectively). Each data set comprised 300 verification–forecast pairs. There is no obvious deviation from reliability, although the histogram for lead time $L = 20$ might be slightly slanted to the right by visual inspection. The test detects no significant deviation from reliability though.

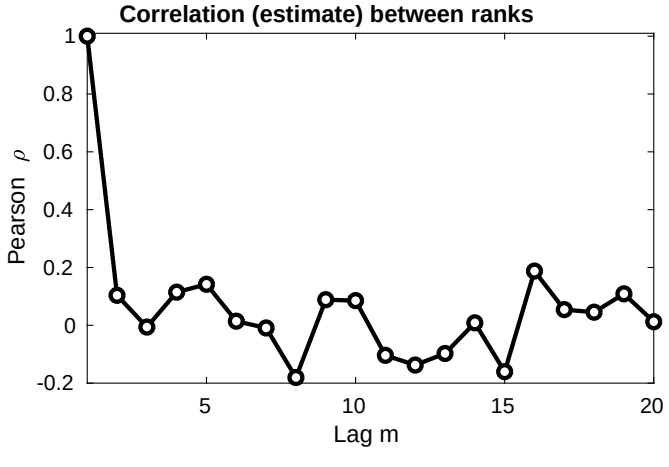

Figure 6. Estimated Pearson correlation coefficient between the ranks R_n and R_{n+m} for ensemble forecasts for the 2D Navier–Stokes. The lead time in this case was 20 time units; values of the lag m between zero and 20 are shown on the abscissa. Although no uncertainty information has been included such as error bars, it is evident that the correlation decreases indeed very quickly with increasing lag, and correlation with larger lag do not contribute much to Υ, due to fast decay of correlation in this system. This implies that the test in this example has properties very similar to the standard GOF test.