Stratified rank histograms for ensemble forecast verification under serial dependence

Article

Accepted Version

Bröcker, J. and Bouallègue, Z. B. (2020) Stratified rank histograms for ensemble forecast verification under serial dependence. Quarterly Journal of the Royal Meteorological Society, 146 (729). pp. 1976-1990. ISSN 0035-9009 doi: https://doi.org/10.1002/qj.3778 Available at http://centaur.reading.ac.uk/90143/

Published version at: http://dx.doi.org/10.1002/qj.3778

To link to this article DOI: http://dx.doi.org/10.1002/qj.3778

Publisher: Royal Meteorological Society

Stratified rank histograms for ensemble forecast verification under serial dependence

Jochen Bröcker[a*], Zied Ben Bouallègue[b]

[a] *School of Mathematical and Physical Sciences, University of Reading, United Kingdom*

[b] *European Centre for Medium Range Weather Forecasts, Reading, United Kingdom*

*Correspondence to: Department of Mathematics and Statistics, Univ. of Reading, Whiteknights, Reading, Berks, RG6 6AX, j.broecker@reading.ac.uk

Rank histograms are a popular way to assess the reliability of ensemble forecasting systems. If the ensemble forecasting system is reliable, the rank histogram should be flat, "up to statistical fluctuations". There are two long noted challenges to this approach. Firstly, uniformity of the overall distribution is implied by but does not imply reliability; ideally the distribution of the ranks should be uniform even *conditionally* on different forecast scenarios. Secondly, the ranks are serially dependent in general, precluding the use of standard goodness–of–fit tests to assess the uniformity of rank distributions without any further precautions. The present paper deals with both these issues by drawing together the concept of stratified rank histograms, which have been developed to deal with the first issue, with ideas that exploit the reliability condition to manage the serial correlations, thus dealing with the second issue. As a result, tests for uniformity of stratified rank histograms are presented that are valid under serial correlations.

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

Received . . .

1. Introduction

To an increasing degree, dynamical forecasting systems for the atmosphere and the ocean are issuing ensemble forecasts, in an attempt to convey a range of future scenarios of the system under concern together with their respective likelihood. There exists by now a large body of work concerned with assessing the quality and skill of ensemble forecasting system, providing both methodological insight as well as practical tools. Several statistical properties of ensemble forecasting systems have been identified as desirable; see for instance Bröcker (2009, 2012); Weigel (2011). An important one is *reliability*, which means (roughly speaking) that at any point t_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. (A formal definition will be given in Section 2.)

Reliability of ensemble forecasts has been considered in a number of publications; a popular tool to assess reliability are rank histograms (see e.g. Anderson 1996; Hamill and Colucci 1997; Talagrand et al. 1997; Hamill 2001). Assuming that the verifications are real numbers, one determines the rank $R(n)$ of the verification $Y(n)$ among the ensemble members $X_1(n), \ldots, X_K(n)$ (where n is the time). If the ensemble forecasting system is reliable, the ranks are uniformly distributed, whence a reliable ensemble forecasting system should produce a "more or less" uniform rank histogram, that is up to random fluctuations.

As has been emphasised by several authors (Hamill and Colucci 1997, 1998; Bröcker 2008; Siegert et al. 2012), uniform rank distribution is only a necessary but not a sufficient criterion for reliability. Potentially more powerful tests result if the verification–forecasts pairs are *stratified*, that is, divided into subsets corresponding to different forecasting situations (roughly speaking). Given reliability, even individual histograms for the separate strata should exhibit a uniform distribution.

Irrespective of whether stratified or unstratified histograms are used, a rigorous testing methodology needs to take into account that a rank histogram will never be precisely flat even for a reliable forecasting system, and the random deviations from flatness need to be analysed quantitatively. As has been noted by several authors (Wilks 2010; Pinson et al. 2010; Siegert et al. 2017; Bröcker 2018), a problem with forecast assessment in general is that the verification–forecast pairs (or in our case the ranks) are generally not independent, which renders this analysis very difficult. In particular, classical goodness–of-fit tests are not applicable to the flatness of rank histograms (stratified or unstratified) since the ranks are serially dependent.

The purpose of the present paper is to extend the approach taken in Bröcker (2018), which addresses this problem in the context of unstratified rank histograms, and extend it to stratified rank histograms. The approach rests on the observation that a reliable ensemble $\mathbf{X}(n) = (X_1(n), \ldots, X_K(n))$ provides (an approximation to) the distribution of $Y(n)$, given what information was available to the forecaster at the initialisation time of the forecast $\mathbf{X}(n)$ (usually at time $n-T$, where T is the lead time). This can be harnessed to at least constrain the correlation structure of the ranks to some extent. The result of the presented analysis is a generalised χ^2–test for the (joint) flatness of stratified rank histograms and thus for the reliability of ensemble forecasts, extending the results in Bröcker (2018).

In Section 2 we present the mathematical setup and provide a definition of reliability for ensemble forecasting systems in mathematical terms. The concept of stratification is explained in Section 3, while Section 4 presents a methodology to statistically test stratified rank histograms for flatness; Theorem 1 provides the asymptotic distribution of the test statistic under reliability and minimal additional assumptions. Section 5 discusses the main steps to perform the test in an algorithmic fasion. Numerical examples are presented in Section 6, discussing data from an operational ensemble forecasting system; Section 7 concludes. Several mathematical details are presented in the Appendices.

2. Setup, notation and the definition of reliability

We start with fixing some general notation. The general setup will be very similar to the one in Bröcker (2018). The verifications are modelled as a sequence $\{Y(n), n = 1, \ldots, N\}$ of random variables with values in the real numbers, with the index n representing the time. The corresponding ensembles $\{\mathbf{X}(n), n = 1, \ldots, N\}$ are modelled as a sequence of random variables, where for each time instant n the ensemble is given by a vector $\mathbf{X}(n) = (X_1(n), \ldots, X_{K-1}(n))$ of $K-1$ ensemble members, where each ensemble member $X_k(n)$ is again a real number.[*] The rank $R(n)$ of the verification $Y(n)$ with respect to the ensemble $\mathbf{X}(n)$ is defined as one plus the number of ensemble members $X_1(n), \ldots, X_{K-1}(n)$ that are smaller than or equal to $Y(n)$.

A desirable property of forecasting systems is *reliability*, which means roughly speaking that for each time n, each individual ensemble member $X_k(n), k = 1, \ldots, K-1$ as well as the corresponding verification $Y(n)$ are drawn independently from the same underlying distribution. To make this precise, for every time instant $n = 1, \ldots, N$ we let $\mathcal{F}_n$ be the information available to the forecaster for producing the ensemble forecast $\mathbf{X}(n)$, that is to say, at the time this ensemble forecast is issued. Further, let

$$p_n(A) := \mathbb{P}(Y(n) \in A | \mathcal{F}_n) \tag{1}$$

be the conditional distribution of $Y(n)$ given the information $\mathcal{F}_n$ for all $n = 1, \ldots, N$ and any set A on the real line.[†] Then the forecasting system is reliable if

$$\begin{aligned} &\mathbb{P}(Y(n) \in A_0, X_1(n) \in A_1, \ldots, X_{K-1}(n) \in A_{K-1} | \mathcal{F}_n) \\ &= p_n(A_0) \cdot \ldots \cdot p_n(A_{K-1}) \end{aligned} \tag{$\mathcal{H}_0$}$$

for all times $n = 1, \ldots, N$ and any selection of subsets $A_0, \ldots, A_{K-1}$ of the real line. The condition ($\mathcal{H}_0$) constitutes the null hypothesis for which tests will be presented. An equivalent formulation is: For all times $n = 1, \ldots, N$,

i. the distribution of each ensemble member $X_k(n)$, conditional on $\mathcal{F}_n$, is equal to the distribution of the verification $Y(n)$, conditional on $\mathcal{F}_n$, and

ii. the ensemble members and the verification $Y(n), X_1(n), \ldots, X_{K-1}(n)$, conditional on $\mathcal{F}_n$, are independent from one another.[‡]

We will impose an additional assumption which is usually not stated as part of the reliability condition but which is evidently satisfied in most applications where forecasts are made with a certain *lead time* T. This means that for any n the forecast $\mathbf{X}(n)$ is prepared a certain number T of time steps previously, implying that at that point the forecaster knows the verifications $Y(m)$ and ensembles $\mathbf{X}(m)$ for $m = 1, \ldots, n-T$. In other words, we assume that

For any $n = 1, 2, \ldots$, the forecast information $\mathcal{F}_n$ contains the verifications and ensembles $Y(m)$ and $\mathbf{X}(m)$ for $m = 1, \ldots, n-T$. (2)

This assumption will be crucial later on. Note however that this assumption does not form part of our null hypothesis as we are not aiming to test against any alternatives to this assumption.

3. Stratification of ensemble forecasts

As we will see below, reliability implies that the ranks $\{R(n), n = 1, 2, \ldots\}$ have a uniform distribution (over the numbers $1, \ldots, K$) but we will see much more, namely that the distribution is uniform *conditionally* on $\mathcal{F}_n$. In broad terms this means that if the entire data set is divided into subsets that correspond to different forecasting scenarios, the ranks within each subset are expected to exhibit a uniform distribution. Dividing the data into subsets that correspond to different forecasting scenarios will be referred to as stratification in the following. The fact that the ranks are uniform within each stratum is a much stronger property than being unconditionally uniform and ought to be exploited for a reliability test. There are various ways to stratify the data, that is, to distinguish between different forecasting scenarios. Here are a few examples:

- If the ensembles are generated by perturbing an analysis (which in turn has been obtained through data assimilation), then that analysis could be used to identify different forecasting situations, and the data could be stratified along the analysis.
- Stratification could be performed directly along observations which are available at forecast time. These could either be observations used to verify previous forecasts, or other observations (of different meteorological quantities for instance).
- The ensemble forecasts could be stratified along another deterministic forecast generated in tandem with the ensemble, such as the high resolution forecast at the European Centre for Medium Range Weather Forecasts.

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

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

[‡] It turns out that the entire analysis in the present paper remains valid if "independence" in this statement is replaced by the weaker condition of "exchangeability" (Bröcker and Kantz 2011).

To describe the idea in mathematical terms, we assume a sequence $\{S(n), n = 1, \ldots, N\}$ of random variables with values in the finite set $\{1, \ldots, L\}$ where $S(n)$ indicates the relevant stratum (out of L different possibilities) at time n. In the examples above, $S(n)$ would be known at forecast time and hence be completely determined by the information in $\mathcal{F}_n$; we will call stratifications of this kind *external stratifications.*

An alternative possibility might come to mind, namely calculating $S(n)$ as a function of the ensemble $\mathbf{X}(n) = (X_1(n), \ldots, X_{K-1}(n))$. Such a function would have to be symmetric as the ordering of the ensemble member does not carry any significant information. This approach has problems though; as was shown in Siegert et al. (2012), stratifying along a symmetric function of the ensemble alone does not give flat rank histograms.

The trick to avoid that difficulty is to include the verification in the stratification function. We consider a symmetric function

$$s : \mathbb{R}^K \to \{1, \ldots, L\} \tag{3}$$

where $L \in \mathbb{N}$, and define the random variables $\{S(n), n = 1, 2, \ldots\}$ through

$$S(n) = s(Y(n), \mathbf{X}(n)) \tag{4}$$

for $n = 1, 2, \ldots$. Possible choices for s are coarse grained versions of the mean or the median. A stratification of this form will be called *internal*. In the following, the stratification might be either external or internal. This will only make a difference with regards to the theory. In practice, external and internal stratifications can be used in exactly the same way.

4. A generalised χ^2–test for flatness of stratified rank histograms

It is now possible to show from Equation ($\mathcal{H}_0$) that for each n fixed, the random variables $R(n)$ and $S(n)$ are independent, and that $R(n)$ has a uniform distribution (see Appendix A). If we denote by $N_{k,l}$ the number of times n for which $R(n) = k$ and $S(n) = l$ where $k = 1, \ldots, K; l = 1, \ldots, L$ and define $N_{\bullet,l} := \sum_{k=1}^{K} N_{k,l}$ (which is the no. of times n for which $S(n) = l$), then by Equation (21) we expect that up to "sampling variations", we have

$$N_{k,l} \cong \frac{1}{K} N_{\bullet,l}. \tag{5}$$

If, in addition, the pairs $(R(n), S(n)), n = 1, 2, \ldots$ were temporally independent, the random variables

$$d_{k,l} = \frac{N_{k,l} - \frac{1}{K} N_{\bullet,l}}{\sqrt{\frac{1}{K} N_{\bullet,l}}}, \tag{6}$$

which basically quantify the error in (5), are asymptotically normal with mean zero and a covariance matrix given by an orthogonal projector onto a $(K-1)L$–dimensional subspace (see for instance Mood et al. 1974, for a discussion of classical χ^2–tests). It follows that the test statistic

$$t = \sum_{k,l} d_{k,l}^2 \tag{7}$$

has a χ–square distribution with $(K-1)L$ degrees of freedom. This fact forms the basis of the classical goodness–of–fit test.

In practice though the pairs $(R(n), S(n)), n = 1, 2, \ldots$ are not temporally independent, but using the reliability condition ($\mathcal{H}_0$) again, now in combination with condition (2) it is possible to obtain strong decorrelation properties of the ranks (see Eq. 23 in Appendix A). It turns out that we also need the rank–stratification pairs $(R(n), S(n)), n = 1, 2, \ldots$ to be a stationary and ergodic sequence. Stationarity of a random sequence $a(1), a(2), \ldots$ means that for any m, the joint distribution of $(a(n), \ldots, a(n+m))$ 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

$$\frac{1}{N} \sum_{n=1}^{N} \phi(a(n), \ldots, a(n+m)) \qquad (m \text{ fixed}) \tag{8}$$

converges to $\mathbb{E}[\phi(a(n), \ldots, a(n+m))]$ 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 is the only extraneous assumption we need to add in order to prove Theorem 1. We note that the rank–stratification pairs $(R(n), S(n)), n = 1, 2, \ldots$ might be a stationary and ergodic sequence even though the original verification–forecast pairs are not. Suppose for instance that the verification–forecast pairs are ergodic "up to" a common deterministic signal $u(n), n = 1, 2, \ldots$ (a climatic trend for instance), in the sense that subtracting this signal from the verification and all ensemble members would render the verification–forecast pairs ergodic. Note that subtracting the signal does not change the ranks, and by choosing a stratification function that does not change either when subtracting the same value from the verification and all ensemble members, we can make sure that the rank–stratification pairs do not depend on this signal and are thus ergodic. For instance if $s(x)$ depends only on differences $x_i - x_j, i, j = 1, \ldots, K$, it will have the required property. The assumption of ergodicity might seem strong, in view of the fact that the relevant data is subject to periodic components (seasonal or diurnal cycles) as well as long term trends such as climate change. A closer analysis reveals that periodic cycles do not present a problem to our methodology if they are much shorter than the overall length of the data set. This is not the case for the seasonal cycle which is one reason why our numerical examples consider data from the winter season only. The only way of dealing with seasonal cycles, it seems, is on a case by case basis. In the same way, there is very little that can be said in general if the data contains fundamental non–stationarities, for instance as a result of climate change (on a time scale comparable to the size of the data archive). It has to be kept in mind though that *any* statistical forecast evaluation method will require some form of stationarity at least, so the concerns here in fact applies to statistical forecast evaluation as a whole. If non–stationarities are present, we lose the link between expected forecast performance in the future and average forecast performance in the past, a link on which statistical forecast evaluation fundamentally rests.

Like the classical goodness–of–fit test, the test proposed here uses a test statistic which will be a modification of t in Eq. 7. Again, the asymptotic distribution of the test statistic will be χ^2 with a certain number of degrees of freedom; this is essentially the statement of Theorem 1 below.

Before stating the theorem, we will try and elucidate the main ideas of the theorem and its proof. It follows from our assumptions that the $d_{k,l}, k = 1, \ldots K, l = 1, \ldots, L$ still satisfy a central limit theorem. In principle, the covariance of the $d_{k,l}$ could be used to normalise these random variables, in order that the sum of their squares again yields a χ^2–distributed quantity. In contrast to the situation with independent ranks though, the asymptotic covariance of these random variables is no longer known. This problem is addressed by estimating the covariance matrix of the $d_{k,l}$ from the data and using this estimate instead of the true covariance matrix. The feasibility of this approach of course requires proof.

The reader might wonder how one might possibly estimate all the required covariances in a real world problem; if we consider for instance an ensemble forecasting system with 50 ensemble members and we want to investiage three strata, the $d_{k,l}$ comprise 153 random variables already, implying in excess of 11,000 covariances to be estimated. In order to reduce that number, we reduce the information taken from each histogram; rather than using the full histogram with its K entries, we project it onto a few elements of $\mathbb{R}^K$ which we call *contrasts*. The effect is that, in statistical terms, the test looses power (the probability of correctly identifying an unreliable forecasting system), but there is better control over the test size (i.e. the significance level is closer to the actual probability of errorneously rejecting a reliable forecasting system as unreliable).

Mathematically speaking, the idea is to choose K–dimensional vectors $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(\mu)}$, the contrasts, and consider the random variables

$$\nu_{m,l} = \sum_{k=1}^{K} \mathbf{w}_k^{(m)} N_{k,l}, \tag{9}$$

for $m = 1, \ldots, \mu$ and $l = 1, \ldots, L$. By taking $\mu < K$ (i.e. fewer contrasts than histogram bars), we obtain a reduction of the dimensionality of the problem. We shall see later that having this option is necessary in practice.

When choosing contrasts, one should avoid "constant" contrasts, that is, contrasts with all components being the same. Indeed, if for instance $\mathbf{w}^{(1)}$ is such a contrast (with all components being one, say), then $\nu_{1,l} = N_{\bullet,l}$ which is simply the number of samples in stratum l and does not contain any information about the histogram. Thus we define a contrast to be a vector $\mathbf{w} \in \mathbb{R}^K$ so that $\sum_{k=1}^{K} \mathbf{w}_k = 0$. Further, we take the contrasts $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(\mu)}$ to be orthogonal and normalised. Such a set can have at most $K-1$ contrasts, so we must have $\mu < K$. Orthogonality of the contrasts would imply that the $\nu_{m,l}$ are asymptotically independent if the rank–stratification pairs were independent (see Jolliffe and Primo 2008, for a thorough analysis in that situation). But even in the general case it seems advisable to use orthogonal contrasts in order that the $\nu_{m,l}$ provide complementary information. A practical way to compute contrasts (with some control on their shape) will be provided in Section 5.1.

Our assumptions will entail that asymptotically (for large N) the quantities

$$\zeta_{m,l} := \frac{\nu_{m,l}}{\sqrt{N}} \tag{10}$$

are jointly (in m, l) normally distributed with mean zero and some covariance tensor $\Upsilon_{m,l,m',l'}$. This covariance will be needed later but is unknown in general, and therefore has to be estimated. An estimator $\hat{\Upsilon}$ will be discussed in Section 5, Equation (17). The feasibility of this estimator is again due to the strong decorrelation property of the ranks (implied by condition (2) and Eq. 20), and the assumption that the pairs $\{(R(n), S(n))\}$ form a stationary and ergodic sequence.

With the inverse $\hat{\Upsilon}^{-1}$ of $\hat{\Upsilon}$ defined so as to satisfy

$$\sum_{m',l'} \hat{\Upsilon}^{-1}_{m,l,m',l'} \hat{\Upsilon}_{m',l',m'',l''} = \delta_{m,m''} \delta_{l,l''}, \tag{11}$$

the proposed test statistic is

$$\tilde{t} := \sum_{m,l,m',l'} \hat{\Upsilon}^{-1}_{m,l,m',l'} \zeta_{m,l} \zeta_{m',l'}, \tag{12}$$

where in the sum the indices m, m' run from 1 to μ, and the indices l, l' run from 1 to L. Using this test statistic is motivated by the fact (already hinted at above) that if the $\zeta_{m,l}$ were indeed normally distributed with mean zero and covariance tensor Υ, then the random variable $\tilde{t}$ in Equation (12) (with Υ in place of $\hat{\Upsilon}$) would have a χ^2–distribution with $\mu \cdot L$ degrees of freedom, as is easily seen. Our theorem states that under the imposed conditions, this is still the case asymptotically for large N.

Theorem 1 *Suppose that the ensemble forecasting system is reliable (i.e. condition $\mathcal{H}_0$ holds), condition* (2) *is satisfied, and $\{(R(n), S(n)), n = 1, 2, \ldots\}$ is ergodic. Then the statistic $\tilde{t}$ has, asymptotically for large N, a χ^2–distribution with $\mu \cdot L$ degrees of freedom.*

For a proof, see Appendix C. By rejecting the hypothesis ($\mathcal{H}_0$) when $\tilde{t} > \theta$ and otherwise accepting, we obtain a test for reliability which according to Theorem 1 is of size $\Phi(\theta)$ (asymptotically for large N), where Φ is the cumulative distribution function of the χ–square distribution with $L \cdot \mu$ degrees of freedom.

Unfortunately, very little of generality can be said about the power of the test. The alternative hypothesis comprises all probability distributions that do *not* satisfy the hypothesis ($\mathcal{H}_0$), and given the multitude of these there is little hope that the presented (or in fact any) test develops nontrivial power against all conceivable alternatives. Furthermore, there does not seem to be an obvious candidate of a restricted alternative hypothesis (or deviation from reliability) that is sufficiently ubiquitous in order to warrant closer investigation and, at the same time, sufficiently specific so as to allow us to make statements about the power. Therefore, as far as we can see a systematic power study would require considering a large number of possibly relevant situations, which is beyond the scope of the present paper.

5. Description of algorithms

In this section, we will list the necessary steps to calculate $\tilde{t}$ and perform the test, although this information could in principle be gathered from Section 3 (with the exception of the estimator for Υ in Equation (17) below). An algorithm to calculate contrasts will also be provided. We still assume that for each n, the verification $Y(n)$ is a real number and the ensemble $\mathbf{X}(n) = (X_1(n), \ldots, X_{K-1}(n))$ is a $K-1$–dimensional vector, that is, an element of $\mathbb{R}^{K-1}$; so there are $K-1$ ensemble members. We let

$$s : \mathbb{R}^K \to \{1, \ldots, L\} \tag{13}$$

be a symmetric function (with values in the set $\{1, \ldots, L\}$). Further, $\{R(n), n = 1, 2, \ldots\}$ are the ranks and $\{S(n), n = 1, 2, \ldots\}$ the strata defined as $S(n) = s(Y(n), \mathbf{X}(n))$ for $n = 1, 2, \ldots$, in case internal stratification is used. Otherwise, let $S(n), n = 1, 2, \ldots$ be indicators of the external strata.

5.1. *Creating a set of contrasts*

We describe an algorithm to create a set $\{\mathbf{w}^{(m)} \in \mathbb{R}^K, m = 1, \ldots \mu\}$ of contrasts, where necessarily $\mu < K$.

I. Let V be a matrix of dimension $K \times (\mu + 1)$ with rank $(\mu + 1)$ (i.e. the columns are linearly independent) and the first column being a constant vector (i.e. all entries are the same and not zero). An example for such a matrix (which gives quite interpretable results) is

$$\mathsf{V}_{k,l} = \left(\frac{k}{K+1} - \frac{1}{2} \right)^{l-1} \tag{14}$$

II. Let Q, R be matrices of dimension $K \times (\mu + 1)$ and $(\mu + 1) \times (\mu + 1)$, respectively, so that

(a) the columns of Q are normalised and mutually orthogonal;

(b) R is right upper triangular;

(c) $\mathsf{V} = \mathsf{QR}$.

Such matrices can be found by applying a Gram–Schmidt procedure to the columns of V or equivalently through a QR–decomposition of V.

III. Now ignore the first column of Q which will have constant entries; the remaining μ columns form the desired contrasts.

Figure 1 shows three contrasts for the case of $K = 8$. These were obtained by applying the described procedure to the matrix in Equation (14) with $K = 8$ and $\mu = 3$.

5.2. *Implementing the generalised χ^2–test of Theorem 1*

We assume that ensembles and verifications have been converted to ranks $\{R(n), n = 1, \ldots, N\}$ and strata $\{S(n), n = 1, \ldots, N\}$. Contrasts $\{\mathbf{w}^{(m)}, m = 1, \ldots, \mu\}$ have also been chosen with $\mu < K$. The lead time is assumed to be T.

I. Calculate $Z_{m,l}(n)$ for $m = 1, \ldots, \mu$, $l = 1, \ldots, L$ and $n = 1, \ldots, N$ according to

$$Z_{m,l}(n) = \delta_{S(n),l} \cdot \mathbf{w}^{(m)}_{R(n)}, \tag{15}$$

where here and in the following we define $\delta_{k,l} = 1$ if $k = l$ and zero otherwise.

II. Calculate $\zeta_{m,l}$ for $m = 1, \ldots, \mu$ and $l = 1, \ldots, L$ according to

$$\zeta_{m,l} = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} Z_{m,l}(n). \tag{16}$$

(Note that this indeed gives the same as Eq. 10.)

III. Estimate the covariance Υ by

$$\begin{aligned}\hat{\Upsilon}_{m,l,m',l'} &:= \frac{1}{N} \sum_{n=1}^{N} Z_{m,l}(n) Z_{m',l'}(n) \\ &+ \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{T-1} \{Z_{m,l}(n) Z_{m',l'}(n+k) + Z_{m,l}(n+k) Z_{m',l'}(n)\}.\end{aligned} \tag{17}$$

Note that $\hat{\Upsilon}$ is by construction symmetric, that is $\hat{\Upsilon}_{m,l,m',l'} = \hat{\Upsilon}_{m',l',m,l}$. Therefore, it is sufficient to calculate $\hat{\Upsilon}_{m,l,m',l'}$ for (m, l) either equal to or larger than (m', l') in lexicographic ordering, that is $(m, l) > (m', l')$ if either $m > m'$ or $m = m', l > l'$.

IV. Find the inverse $\hat{\Upsilon}^{-1}$ of $\hat{\Upsilon}$ (in the sense of Eq. 11) and calculate the test statistic

$$\tilde{t} = \sum_{m,l,m',l'} \hat{\Upsilon}^{-1}_{l,m,l',m'} \, \zeta_{m,l} \, \zeta_{m',l'}. \tag{18}$$

V. Compare $\tilde{t}$ to a χ–square distribution with $L \cdot \mu$ degrees of freedom. That is, let Φ be the cumulative distribution function of the χ–square distribution with $L \cdot \mu$ degrees of freedom, then the p–value of our data is given by $p = 1 - \Phi(\tilde{t})$.

A python package `franz` (Bröcker 2019) has been implemented which provides the described reliability tests as well as methods for computing contrasts. In addition, `franz` contains tests for reliability of other types of forecasts.

6. Numerical Experiments

In this section, we aim to demonstrate how stratified rank histograms can help diagnosing conditional biases and assessing reliability. The examples below are meant to illustrate the interpretation of different shapes of histograms and the use of different types of stratification. We will be using forecasts from the European Centre for Medium Range Weather Forecasts; however, this study should not be considered as a comprehensive analysis of the reliability for this forecasting system for 2 m temperature.

Results for both stratified as well as unstratified tests will be reported. We stress that there is not necessarily a strict relation between the p–values for these tests. When applied to exactly the same data, the p–values for stratified tests might be either higher or lower than for unstratified tests. This might seem odd since a flat unstratified histogram represents a weaker form of reliability than a flat set of stratified histograms. It has to be kept in mind though that the stratified test requires the estimation of more parameters and thus more data might be needed until there is significant evidence to reject the null hypothesis.

6.1. *Data*

Ensemble forecasts of 2 m temperature serve as a basis for the illustration of the methodology and concepts presented in this paper. The dataset comprises observations from SYNOP stations and the corresponding nearest grid point forecasts from the operational ensemble prediction system (ENS) based on the Integrated Forecast System (IFS) of the European Centre for Medium-Range Weather Forecasts. The focus in on a forecast horizon of 5 days, with forecasts valid once per day at 12 UTC. This means that we are working with a lead time of $T = 5$. The ensemble comprises 50 members. The assessment of the ensemble forecast is performed separately for different locations distributed over the European continent . Fig. 2 shows the six SYNOP stations selected for this exercise: Salla in Finland (i), Sankt Peter-Ording in Germany (ii), Cork in Ireland (iii), Beauvais in France (iv), Slatina in Romania (v), and Monte Real in Portugal (vi). We consider four consecutive winters (Dec. 2015–Feb. 2016 to Dec. 2018–Feb. 2019) in order to have consistent datasets in terms of weather conditions as well as samples of reasonable sizes (with 361 measurements at each location, including reported missing values). As a pre-processing step, forecasts are adjusted first by applying an orographic correction that accounts for systematic mismatch between station height and the orography in the model. This adjustment ΔT is linear with the height difference Δz between station and model representation and given by $\Delta T = -0.0065\,\mathrm{K\,m^{-1}}\,\Delta z$. Secondly, raw forecasts provide information on a grid whereas observations are point measurements. This scale mismatch leads to representativeness error in the forecast that are easy to correct for in a simplistic way. The raw ensemble spread, associated with a forecast valid at the model-resolution scale (~ 18 km), can be inflated in order to capture the temperature uncertainty at smaller spatial scale. The method followed here consists in adding to each member a draw from a centred Gaussian distribution with standard deviation

$$\sigma_{pert} := 0.4 + 0.3 \left|\Delta_e\right|^{1/4}, \tag{19}$$

where Δ_e is the altitude difference between station and model representation. This formula is derived from the analysis of 2 m temperature measurements of a high-density observation network over Europe following the same methodology as in Ben Bouallègue et al. (2020).

Bias correction and spread correction are both applied to correct for representativeness error in the observations. So it does not aim at providing a reliable forecast but rather at making a fairer comparison between forecasts and verifications. The model in Eq (19) is valid for forecast on a grid with a resolution of $18\,km$. No further pre– or postprocessing was applied to either forecasts or verifications.

6.2. *Experiments*

Results are shown for the six selected stations with location as shown in Fig. 2. Figs. 3 to 8 correspond to these six locations, respectively. The panels in each figure show the stratified rank histogram (top panel), an unstratified rank histogram for the complete dataset (middle panel) and the corresponding covariance matrix Υ (bottom panel). For illustrative purposes, we applied two different types of stratification: an internal stratification based on the mean over all members and observations, and an external using the 10 m wind forecast valid at the verification time. Both strata were tested for each station but only one will be presented here for illustration purposes. Each stratification subdivides the observation–forecast pairs into three strata, with each stratum containing about a third of all instances. Further, two orthogonal contrasts were used , generated as in Section 5.1, Equation (14). These look basically as the linear and U–shaped contrasts in Figure 1, except that the linear contrasts is decreasing rather than increasing. With regards to choosing the number of contrasts μ and the number of strata L, it needs to be kept in mind that the size of the covariance Υ is $(L \cdot \mu)^2$, and thus the number of parameters to be estimated is roughly $(L \cdot \mu)^2/2$ as the covariance matrix is symmetric. We have N data points but there is dependency among them. It follows from the previous discussion however that N/T can be used as rough estimate for the effective sample size; this is clearly a very pessimistic estimate as it assumes we throw away a fraction $\frac{T-1}{T}$ of the data. We thus arrive at $\frac{TL^2\mu^2}{2N}$ as a rough estimate for the relative error in the estimator for the covariance matrix. In our experiments, we have $N = 361$ and $T = 5$ and we choose $L = 2$ and $\mu = 3$, which gives an error of about 25%. This might seem large but keeping in mind that this is a very pessimistic estimate, we decided this to be acceptable. For the unstratified histograms we have used two contrasts as well for comparison, even though according to the previous considerations there are fewer covariance parameters to be estimated for unstratified histograms so in principle, more contrasts could be used.

The correlation Υ is shown in the third panel, with the field with coordinates $(c_1 + 2(s_1 - 1), c_2 + 2(s_2 - 1))$ corresponding to the entry $\Upsilon_{c_1,s_1,c_2,s_2}$. The sample size as well as the p-value of the reliability statistical test (to four decimal places) are indicated on the top of the plots in each case. Along with the stratified rank histograms, we also indicate the mean value of the stratum for each of the three categories.

(i) Salla

Results for Salla are presented in Fig. 3. The L-shape of the rank histogram indicates that the forecast is positively biased (Fig. 3.b). Stratification based on the mean forecast and observed temperature reveals a conditional bias: over-forecasting occurs only in low temperature conditions (Fig. 3.a) Not surprisingly, the p–values of the reliability tests before as well as after stratification are close to zero. We also see a particularly strong correlation between the two contrasts in the histograms corresponding to low temperature (bottom left in Fig. 3.c). This correlation can be explained directly with the shape of the histogram. We are using a linearly decreasing contrast and a U–shaped contrast; multiplying these gives positive values if the value of the rank is small, and negative values if the rank is large. As the histogram is tilted to the left, small values of the rank are more numerous which implies that the correlation sum is dominated by positive terms, resulting in a positive correlation. Even though here the correlations are estimated not by a complete double sum but by a sum over pairs up to temporal lag L (see Eq. 17), we still expect to see that effect.

(ii) Sankt Peter–Ording

In Fig. 4.b, the unstratified rank histogram is noisy but appears overall quite flat. The reliability test is passed with a p-value of 7%. The test is also successful after stratification with a p-value greater than 10% in that case. Results in Fig. 4.a and 4.c are based on an internal stratification. When a wind-based stratification is applied, the p–values of the reliablity test is close to 30% (not shown). In Fig. 4.c, as expected, the covariance matrix is mainly dominated by the diagonal terms.

(iii) Cork

In Fig. 5, the inverted–U–shape of the rank histogram indicates over-dispersiveness of the ensemble forecast at day 5 for this station. Reliability test fails both in the stratified and unstratified cases. Over-dispersiveness is not a common characteristic of ensemble forecasts for weather surface variables. The interpretation of these results could point to a model deficiency or could lead to question the post-processing step described above. Aiming at accounting for representativeness uncertainty, the spread correction in Eq. (19) is based on the analysis of temperature spatial variability over different seasons and many stations over Europe. So the model is probably too simplistic to describe accurately representativeness uncertainty over winter months at Cork station. But is representativeness uncertainty over-estimated in that case or is the ensemble anyway overdispersive at day 5 for this location and time of the year? The stratification histograms in Fig. 5.a tends to indicate that over-dispersiveness is mainly associated with warm conditions, so related to model limitations. This conclusion is supported by a maximum value in the covariance matrix (Fig. 5.c) reached for warm temperature conditions and the 2nd contrast corresponding to the U-shape (top right corner).

(iv) Beauvais

In Fig. 6, reliability of 2 m temperature ensemble forecasts at Beauvais is investigated. Focusing on the unstratified results, the rank histogram appears flat and the reliability test is successfully passed with a p-value around 23%. However, reliability is rejected under stratification, with the p-value being close to 1%. This is not the case when 10 m wind forecasts are used as a stratum: the test is passed with a p-value of 15% (not shown). In Fig. 6, stratified rank histograms based on the mean forecast and observed temperature reveal that the forecast could suffer from a conditional bias: a negative bias in warm-temperature conditions. This finding is corroborated by the analysis of the covariance matrix in Fig. 6.c which shows an anticorrelation between the histograms with the two contrasts for the last stratification category. Again, this is easily explained given the shape of the histogram as in the

Salla example, except that now the histogram is tilted to the right. High values of the rank are now more numerous so that the U–shaped contrast gives positive values while the linear contrast tends to give negative values. This implies that the correlation sum between them is dominated by negative contributions, resulting in a negative correlation.

(v) Slatina

In Fig. 7.b, the histogram has a U-shape typical of under-dispersive ensemble forecasts. Stratification is this time based on 10 m wind speed forecast at day 5. The reliability tests fail both in the stratified and unstratified cases. In Fig. 7.a, a negative bias dominates the shape of the rank histogram when focusing on low wind conditions (top panel). Conversely, a slight positive bias seems associated with intermediate to high wind conditions. In Fig. 7.c, anti-correlation between histograms with the two different contrasts is more important for the low wind condition category. So the related negative bias could be seen as the main forecat issue for this location.

(vi) Monte Real

The shape of the histogram in Fig. 8.b can be described as a half inverted–U–shape. The larger population for higher ranks indicates the tendency of a negative bias in the ensemble forecast. While the positive bias in Salla is sharp, the negative bias appears here more gradual and diffuse. Stratification is performed using 10 m wind forecasts and shows that under-forecasting affects the ensemble for all wind conditions. Reliability tests fail with p–values below 1% in both cases. Similarly, internal stratification based on the mean temperature does not provide further indications about which weather conditions could favour the forecast bias. The covariance matrix in Fig. 8.c looks also more complex than in the previous examples. Further diagnostic of the ensemble reliability at that location could be performed using different, potentially more informative stratification.

7. Conclusions and outlook

The rank histogram, a widely used tool to assess the reliability of ensemble forecasting systems, was revisited. The rigorous statistical interpretation of rank histograms suffers from two long noted problems, which have been addressed in this work. Firstly, even for a completely reliable forecasting system, the rank histogram will show statistical deviations from flatness, but for a quantitative assessment the distribution of these fluctuations is required (at least asymptotically). Analysing this distribution is rendered difficult by the fact that the ranks, in general, are not independent but exhibit serial correlations. Secondly, uniformity of the overall distribution is necessary but not sufficient for reliability; ideally the distribution of the ranks should be uniform *conditionally* on different forecast scenarios.

The present paper deals with both these issues successfully under conditions that are arguably satisfied in a wide range of applications. The proposed test effectively performs a generalised goodness–of-fit statistic jointly for a set of histograms, each of which represents a subset of the data, referred to as a stratum. Stratification may be performed either along an external variable or along criteria which involve the ensemble and the verification in a suitable way.

The asymptotic distribution of the test statistic is derived rigorously under the null hypothesis plus minimal additional assumptions; firstly, the sequence of verification–forecast pairs needs to be ergodic, and secondly, past verification–forecast pairs need to be available to the forecaster with a certain temporal lag T which we refer to as the lead time. Under these circumstances the ranks will show temporal dependence but only up to T time steps into the past, an observation which turns out to be crucial for our analysis.

Six data sets were analysed using the methodology presented. Each data set comprises 2 m temperature forecasts from the operational ensemble prediction system of the European Centre for Medium Range Weather Forecasts for certain stations over Europe as well as the corresponding verifications. For all of these stations, the stratified rank histograms and the associated tests reveal interesting diagnostic detail which is not available from the unstratified histograms. In the case of Beauvais and Slatina (*iv* and *v*), we see conditional biases in the stratified histograms that get confounded in the unstratified histograms, to the extent that the forecasting system appears to be underdispersive in the case of Slatina or even reliable in the case of Beauvais. For the presented examples the stratified tests will reject the null if the unstratified tests do, implying that there is no indication of the stratified test loosing power. We note that in the case of Cork and Monte Real, (*iii* and *vi*), the p–values for the stratified tests are higher than for the unstratified ones, but all of these numbers are very small and far away from any meaningful significance level. Furthermore, in the case of Salla and Cork (*i* and *iii*), the defects visible in the unstratified histogram seem to originate in a single stratum. For Salla the unstratified histogram suggest a warm bias of the forecasts while the stratified histogram indicates that this bias appears only under cold conditions; for Cork the underdispersiveness of the unstratified histogram seems in fact restricted to warm conditions, only.

We stress, however, that the question as to which stratum or strata cause a rejection of reliability is difficult to answer or even pose meaningfully. Each stratum could be tested individually by simply discarding all instances of the data that are not in that stratum. The interpretation though is hampered by the fact that the strata are not independent and it is therefore difficult to adjust for multiple testing. This is an inevitable consequence of the more complex dependence structure of the problem. Answering under which stratum reliability fails might be possible if the covariance shows a clear block structure as then the strata contribute independently to the statistic; we leave this as a problem for future research.

8. Acknowledgements

Fruitful discussions with Tobias Kuna, Chris Ferro, Thomas Haiden, and Martin Leutbecher are gratefully acknowledged.

A. An important identity regarding the distribution of $R(n)$ and $S(n)$

In this appendix, we will show that for any $k = 1, \ldots, K$ and any $n = 1, \ldots, N$ we have

$$\mathbb{P}(R(n) = k | S(n), \mathcal{F}_n) = \frac{1}{K}. \tag{20}$$

This implies that

$$\mathbb{P}(R(n) = k | S(n)) = \frac{1}{K}, \tag{21}$$

meaning that for each n fixed, the random variables $R(n)$ and $S(n)$ are independent and that $R(n)$ has a uniform distribution.

We introduce the shorthand $\bar{\mathbf{X}}(n) = (Y(n), \mathbf{X}(n))$ and note that reliability (i.e. Eq. $\mathcal{H}_0$) implies that the distribution of $\bar{\mathbf{X}}(n)$, conditionally on $\mathcal{F}_n$ is symmetric. To prove Equation (20), it is sufficient to show that this distribution remains symmetric if $S(n)$ is included in the conditions. (If external stratification is used, then $S(n)$ is part of $\mathcal{F}_n$ by definition so there is nothing to show.) We recall that the function $s : \mathbb{R}^K \to \{1, \ldots, L\}$ which defines the stratification is symmetric. Let π be an arbitrary permutation of K elements, $A \in \mathcal{F}_n$, and $B \subset \mathbb{R}^K$ a measurable set. Then we have

$$\begin{aligned} &\mathbb{P}(\{\bar{\mathbf{X}}(n) \in B\} \cap \{s(\bar{\mathbf{X}}(n)) = l\} \cap A) \\ &= \mathbb{P}(\{\pi(\bar{\mathbf{X}}(n)) \in B\} \cap \{s \circ \pi(\bar{\mathbf{X}}(n)) = l\} \cap A) \\ &= \mathbb{P}(\{\pi(\bar{\mathbf{X}}(n)) \in B\} \cap \{s(\bar{\mathbf{X}}(n)) = l\} \cap A), \end{aligned} \tag{22}$$

where the first equality is due to the distribution of $\mathbf{X}(n)$ being symmetric conditionally on $\mathcal{F}_n$, and the second due to s being symmetric. By standard probability calculus, Equation (22) implies $\mathbb{P}(\{\bar{\mathbf{X}}(n) \in B\} | S(n), \mathcal{F}_n) = \mathbb{P}(\{\pi(\bar{\mathbf{X}}(n)) \in B\} | S(n), \mathcal{F}_n)$, which means that the distribution of $\bar{\mathbf{X}}(n)$, conditionally on $\mathcal{F}_n$ *and* $S(n)$, is symmetric. This implies Equation (20). Equation (21) follows from Equation (20) and the tower property of the conditional expectation.

An important consequence of Equation (20) emerges in combination with condition (2). Taking the expectation of Equation (20) conditionally on $S(n)$ and $\{(R(m), S(m)), m = 1, \ldots, n - T\}$, we can invoke the tower property (thanks to condition (2)) and obtain

$$\mathbb{P}(R(n) = k | S(n), \{(R(m), S(m)), m = 1, \ldots, n - T\}) = \frac{1}{K}. \tag{23}$$

This relation will be important later on.

B. Covariance estimator

In this appendix, we discuss an estimator for Υ, the covariance matrix of $\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbf{Z}(n)$ in the limit $N \to \infty$, where $\mathbf{Z}(n) = (Z_{m,l}(n))_{m,l}$; otherwise, notation and definitions are as in Sec. 2. We start with defining the (matrix valued) covariance function

$$\gamma(k) := \mathbb{E}(\mathbf{Z}(n)\mathbf{Z}(n+k)^T), \tag{24}$$

noting that since $\{\mathbf{Z}(n), n = 1, 2, \ldots\}$ is stationary there is no dependence on n. Furthermore, γ is well defined for negative k, too, and in fact $\gamma(-k) = \gamma(k)^T$. An elementary calculation then gives

$$\Upsilon = \sum_{k \in \mathbb{Z}} \gamma(k), \tag{25}$$

provided the sum converges. But thanks to Equation (23), we have $\gamma(l) = 0$ if $l \geq T$, meaning that the sum in Equation (25) contains only finitely many nonzero terms, namely for $|k| < T$. These terms can be estimated by empirical averages (i.e. averages over time), that is

$$\gamma_N(k) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{Z}(n)\mathbf{Z}(n+k)^T, \tag{26}$$

which converges to $\gamma(k)$ for $N \to \infty$, due to the condition that $\{(R(n), S(n))\}$ are ergodic. By replacing $\gamma(k)$ in Equation (25) with the estimators $\gamma_N(k)$, we obtain the estimator $\hat{\Upsilon}$ for Υ given in Equation (17).

C. Proof of the theorem (sketch)

In this appendix, we justify 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 $\delta_N := \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \Lambda(n)$ 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 appendices entail that $\{\Lambda(n), n = 1, 2, \ldots\}$ is a stationary and ergodic process with the property that if $k \geq T$ and $n \geq m$, then

$$\mathbb{E}(\Lambda(n+k) | \Lambda(n), \ldots, \Lambda(m)) = 0. \tag{27}$$

It can be shown that the process $\{\Lambda(n)\}$ can be extended to negative times, and that Equation (27) still holds in the limit $m \to -\infty$. As a result, the conditions of Theorem 4.18 in van der Vaart (2010) are satisfied and we can conclude that the distribution of δ_N 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.

Zied Ben Bouallègue, Thomas Haiden, Nicholas J. Weber, Thomas M. Hamill, and David S. Richardson. Accounting for representativeness in the verification of ensemble precipitation forecasts. 2020. (submitted).

Jochen Bröcker. On reliability analysis of multi-categorical forecasts. *Nonlinear Processes in Geophysics*, 15(4):661–673, 2008. ISSN 1023-5809. URL http://www.nonlin-processes-geophys.net/15/661/2008/.

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. Assessing the reliability of ensemble forecasting systems under serial dependence. *Quarterly Journal of the Royal Meteorological Society*, 2018. doi: 10.1002/qj.3379. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3379. (accepted).

Jochen Bröcker. `franz`; a python library for statistical assessment of forecasts. *GitHub repository*, 2019. URL https://github.com/eirikbloodaxe/franz.

Jochen Bröcker and Holger Kantz. The concept of exchangeability in ensemble forecasting. *Nonlinear Processes in Geophysics*, 18 (1):1–5, 2011. doi: 10.5194/npg-18-1-2011.

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.

Thomas M. Hamill and Stephen J. Colucci. Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Monthly Weather Review*, 126:711–724, March 1998.

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.

Alexander M. Mood, Franklin A. Graybill, and Duane C. Boes. *Introduction to the Theory of Statistics.* McGraw-Hill Series in Probability and Statistics. McGraw-Hill, 1974.

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.

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. 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`.

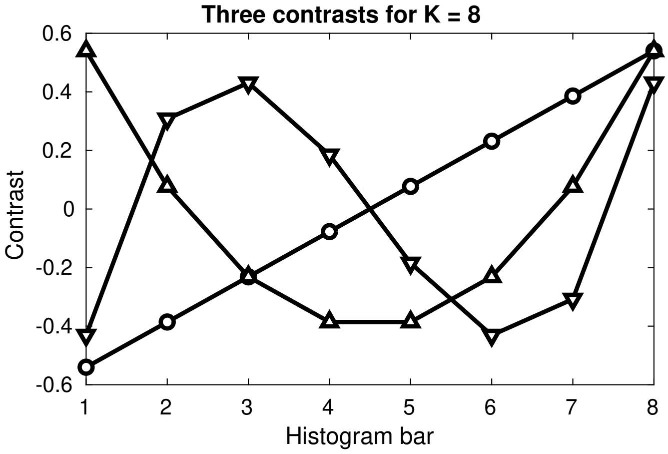

Figure 1. The figure shows three contrasts for the case of $K = 8$. These were obtained by applying the procedure described in Section 5.1 to the matrix in Equation (14) with $K = 8$ and $\mu = 3$. (Lines connecting the points are merely for guidance.)

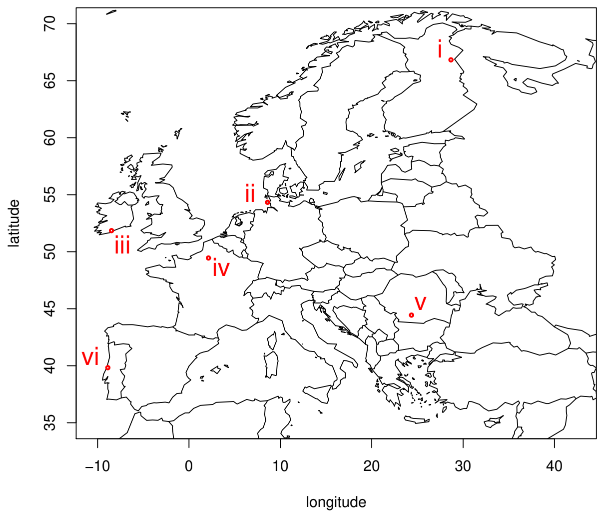

Figure 2. Distribution of the selected stations, i: Salla (Finland), ii: Sankt Peter-Ording (Germany), iii: Cork (Ireland), iv: Beauvais (France), v: Slatina (Romania), vi: Monte Real (Portugal)

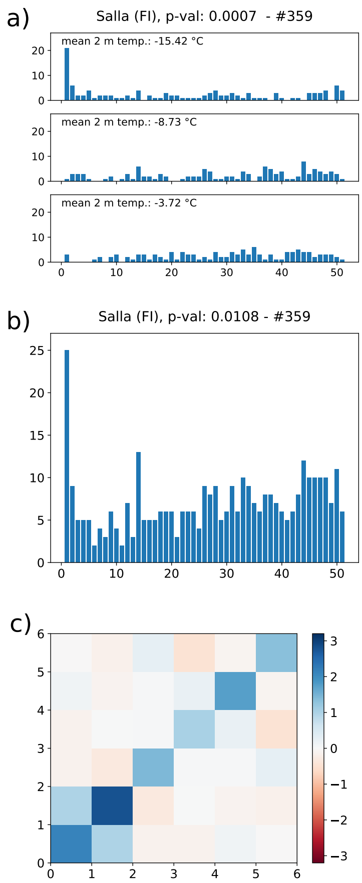

Figure 3. Stratified rank histogram (a), rank histogram (b), and corresponding covariance matrix Υ (c) for Salla (Finnland). Stratification is based on averaged forecast and observed 2 m temperature. The average of this quantity across the stratum is indicated in the corresponding sub–panel of (a). The p–value of the reliability test as well as the sample size (number of forecast-observation pairs) are indicated above the left and middle panels. The unstratified histogram shows a warm forecast bias; The stratified histogram indicates that this is confined to cold conditions.

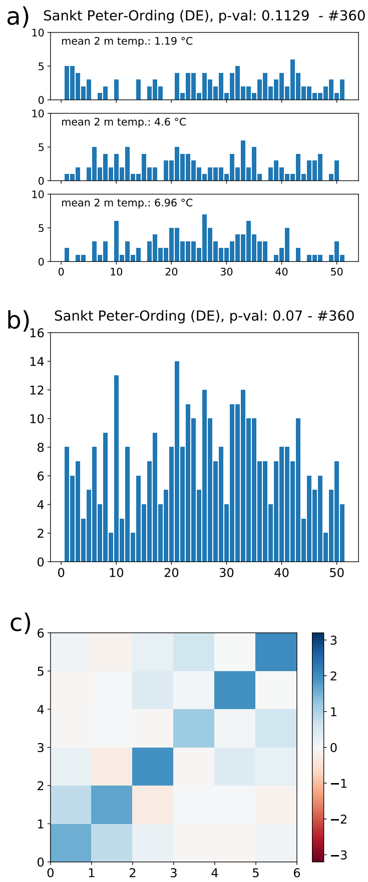

Figure 4. Same as Fig. 3 but for Sankt Peter-Ording (DE). There is no clear indication to reject reliability.

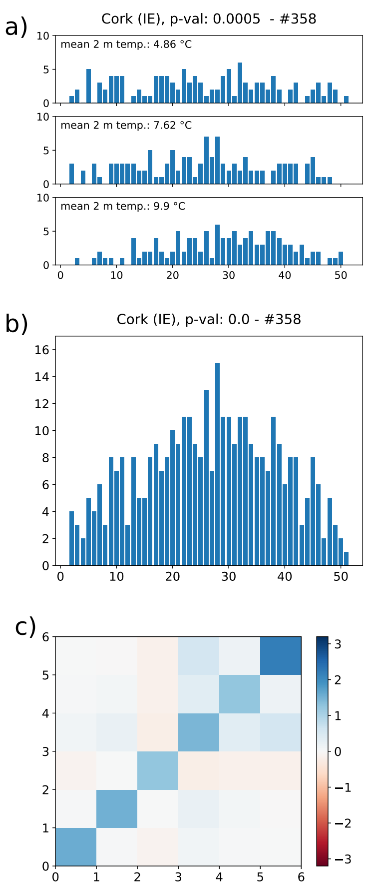

Figure 5. Same as Fig. 3 but for Cork (IE). The forecast overdispersion shown by the unstratified histogram appears to be genuine but probably confined to warm conditions.

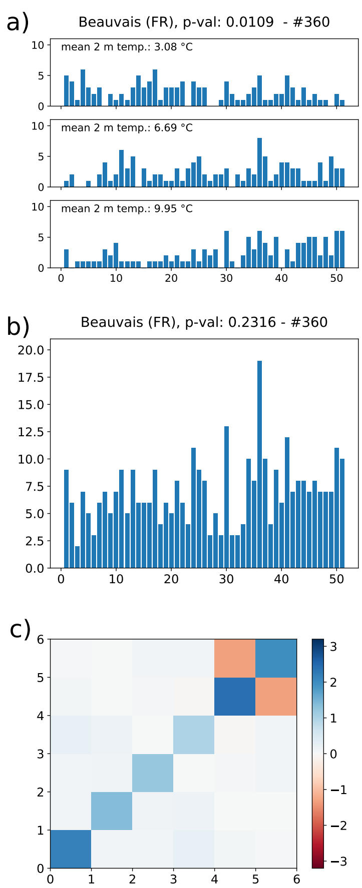

Figure 6. Same as Fig. 3 but for Beauvais (FR). The unstratified analysis provides no evidence to reject reliability, but the stratified histogram indicates conditional bias in different directions under cold vs warm conditions, and therefore evidence to reject reliability.

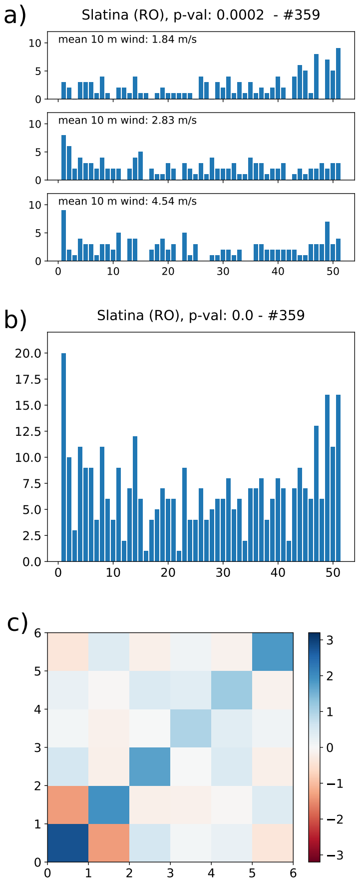

Figure 7. Same as Fig. 3 but for Slatina (RO). Stratification is based on forecast 10 m wind speed. This time the unstratified analysis indicates a lack of spread, but as for Beauvais, the stratified histogram indicates conditional bias in different directions under cold vs warm conditions as evidence to reject reliability, rather than problems with spread.

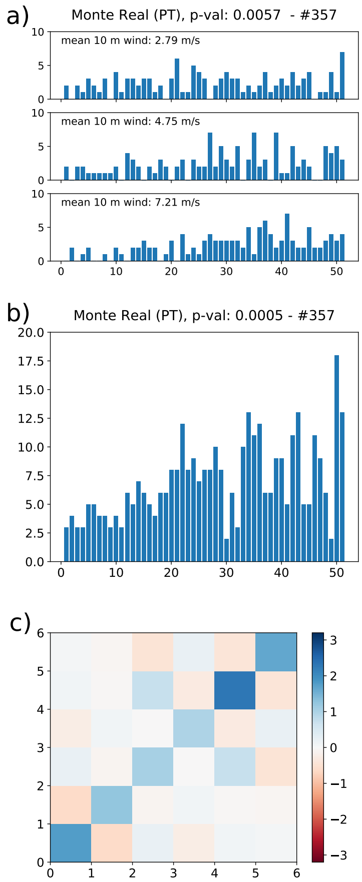

Figure 8. Same as Fig. 7 but for Monte Real (PT). The unstratified histogram shows a cold bias of the forecast which is also present under stratification in all conditions.