Testing the reliability of forecasting systems

Journal of Applied Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

Testing the reliability of forecasting systems

J. Bröcker

To cite this article: J. Bröcker (2021): Testing the reliability of forecasting systems, Journal of Applied Statistics, DOI: 10.1080/02664763.2021.1981833

To link to this article: https://doi.org/10.1080/02664763.2021.1981833

JOURNAL OF APPLIED STATISTICS
https://doi.org/10.1080/02664763.2021.1981833

Taylor & Francis
Taylor & Francis Group

Testing the reliability of forecasting systems

J. Bröcker

School of Mathematical, Physical and Computational Sciences, University of Reading, Reading, UK

ABSTRACT

The problem of statistically evaluating forecasting systems is revisited. The forecaster claims the forecasts to exhibit a certain nominal statistical behaviour; for instance, the forecasts provide the expected value (or certain quantiles) of the verification, conditional on the information available at forecast time. Forecasting systems that indeed exhibit the nominal behaviour are referred to as *reliable*. Statistical tests for reliability are presented (based on an archive of verification–forecast pairs). As noted previously, devising such tests is encumbered by the fact that the dependence structure of the verification–forecast pairs is not known in general. Ignoring this dependence though might lead to incorrect tests and too-frequent rejection of forecasting systems that are actually reliable. On the other hand, reliability typically implies that the forecast provides information about the dependence structure, and using this in conjunction with judicious choices of the test statistic, rigorous results on the asymptotic distribution of the test statistic are obtained. These results are used to test for reliability under minimal additional assumptions on the statistical properties of the verification–forecast pairs. Applications to environmental forecasts are discussed. A python implementation of the discussed methods is available online.

ARTICLE HISTORY
Received 8 January 2020
Accepted 14 September 2021

KEYWORDS
Forecasting; reliability; identifiability; environmental statistics

1. Introduction

There is by now a vast literature on statistical evaluation of forecasts, and a large variety of tools and performance indices have been devised, depending on the nature of the forecasts (probabilities, ensembles, moments, intervals, etc.), the application (cost–loss ratios, decision scenarios, economic value, etc.) and the nature of the verification (binary, vector-valued, spacial fields, etc). Regarding the evaluation of probability forecasts, a classical article is [10]. The meteorological community has contributed significantly to forecast evaluation, both out of academic interest but also due to the societal need for accurate forecasts of meteorological phenomena. Various industries, as well as public and private sector institutions, are reliant on meteorological forecasts to operate successfully. Hence there is a need to evaluate the performance of forecasting systems in an objective manner, either

CONTACT J. Bröcker j.broecker@reading.ac.uk School of Mathematical, Physical and Computational Sciences, University of Reading, Reading, RG6 6AX UK

Supplemental data for this article can be accessed here. https://doi.org/10.1080/02664763.2021.1981833

with regard to certain specific applications or in a general sense. For an overview (in a meteorological context), see the book [44] or the collection [24], containing chapters on ensemble forecasts [42] and on probability forecasts [7]. Evaluation of meteorological forecasts with a focus on applications in economics and decision making has been considered in [25,27,32]; extreme event forecasting has received special attention, see for instance [16,39]. Forecasts for binary events are considered in [23,30]. For the case of several categories, see for instance [2].

In the present paper, rather than quantifying the 'accuracy' of forecasting systems as in [10,21] and a large body of subsequent research or comparing several forecasting systems in terms of performance [13,17,29], we are interested in assessing whether a forecasting system adheres to a desired nominal behaviour. Examples for desired nominal behaviour could be that the forecasting system produces the expected value of the verification (conditional on the information available to the forecaster at forecast time), or as another example, that the forecasting system produces ensembles that are independent draws from the distribution of the verification (again, conditional on the available information at forecast time). Forecasting systems that adhere to the desired nominal behaviour will also be referred to as *reliable*. (The term *calibration* is sometimes used synonymously, although mostly in connection with probability forecasts, i.e. forecasts that under nominal behaviour represent the entire conditional distribution of the verification.)

Our tests for reliability will use an archive of verification–forecast pairs $\{(Y(k), f(k)), k = 1, \ldots, n\}$. Unless otherwise stated, the temporal index (mostly appearing as an argument in round brackets) refers to the *verification time*, that is $f(k)$ is the forecast for the verification $Y(k)$ which becomes available at time k. In most applications, this forecast will have to be issued at some point of time prior to k. The tests will evaluate each forecast $f(k)$ against the corresponding verification $Y(k)$ by means of an identification function Φ which takes both $Y(k)$ and $f(k)$ as arguments and has values in $\mathbb{R}^D$ for some D (which might be larger than one). We use the shorthand $\phi(k) := \Phi(Y(k), f(k))$. The identification function, we assume, is chosen so that if the forecasts adhere to nominal behaviour, the quantity $\phi(k)$ has zero expectation, conditional on the information available to the forecaster at the time when she has to issue the forecast $f(k)$. Thus the identification function will help to quantify (roughly speaking) the deviation from nominal behaviour. Two examples (which we will revisit and treat more formally in Section 2) shall illustrate these concepts.

Example 1.1 (Conditional mean forecasts): Suppose both verifications and forecasts are real numbers. The forecaster claims that for all k we have

$$\mathbb{E}(Y(k) \mid \mathcal{F}(k)) = f(k), \tag{1}$$

where $\mathcal{F}(k)$ denotes the information available at forecast time (later formalised as a filtration). Since the forecast $f(k)$ itself is available at forecast time, Equation (1) is equivalent to $\mathbb{E}(\phi(k) \mid \mathcal{F}(k)) = 0$ for all k, provided we take $\Phi(y, f) := y - f$ as identification function. The image of this identification function is one-dimensional, that is $D = 1$.

Example 1.2 (Probability forecasts for binary events): In this example, forecasts are real numbers between 0 and 1, while verifications assume the values 0 or 1, only. The forecaster claims that for all k we have $\mathbb{P}(Y(k) = 1 \mid \mathcal{F}(k)) = f(k)$, where again $\mathcal{F}(k)$ denotes the information available at forecast time. In other words, a reliable $f(k)$ agrees with

the conditional probability of the event $Y(k) = 1$, given that information. This time, we may take the (one-dimensional) identification function $\Phi(y,f) := \frac{y-f}{\sqrt{f(1-f)}}$. We obtain that not only $\mathbb{E}(\phi(k) \mid \mathcal{F}(k)) = 0$ but also $\mathbb{E}(\phi(k)^2 \mid \mathcal{F}(k)) = 1$, provided the forecasts are reliable. We will call such identification functions *standardised* (see Section 2). In Example 2.2 (Section 2), we will generalise the present situation to verifications with more than two categories. The corresponding forecasts will then be probability vectors in a higher-dimensional space and also the image of the identification function will have to be higher-dimensional.

Our general approach will be to consider

$$S(n) = \sum_{k=1}^{n} \phi(k). \tag{2}$$

(We will later introduce somewhat more general test statistics involving stratification, which give a more detailed picture of reliability, see Section 2.2.) Choosing an identification function that is expected to be zero if the forecasting system is indeed reliable, we would expect $S(n)/n$ to be small. This will be made precise later; the key technical issue in our framework then becomes to understand the distribution of $S(n)/n$ (at least asymptotically) under the null hypothesis of reliability.

The difficulty lies in the fact that the statistical properties of the time series $\{\phi(k), k \in \mathbb{N}\}$ might be very complicated, and as several authors have noted (for instance [36,43]), the verification–forecast pairs can certainly not assumed to be independent. Even if the application of the Law of Large Numbers can be justified, the deviations of $S(n)/n$ from zero need to be quantified, which calls for a Central Limit Theorem. And granted the Central Limit Theorem, there remains the problem of determining the proper scaling or variance. As [43] has pointed out, the assumption of independence can lead to a serious underestimation of the variance and thus overly optimistic (i.e. too narrow) confidence intervals. In the same paper, the evaluation of probability forecasts is investigated using explicit (parametric) assumptions regarding the dependence structure and distribution of the forecasts, but the considered situation is very specific.

On the other hand, some information about the dependence structure in the verification is available through the forecasts. Assuming reliability, $f(k)$ provides information about $Y(k)$, given the information available at the time the forecast $f(k)$ was issued. This information may be harnessed to provide (at least partly) the correlation structure of the time series $\{\phi(k), k \in \mathbb{N}\}$. As we will see, this is easiest in the case of forecasting systems predicting a single time step ahead as then $S(n)$ turns out to be a martingale. In case of larger lead times, the situation is more complicated but we will still find that the time series $\{\phi(k), k \in \mathbb{N}\}$ has finite correlation length (see Equation (13) in Section 3.3 for precise statement).

This fact plays an important role in previous work on the statistical evaluation of forecasts, which we will now review briefly. The classical paper [13] proposes a general methodology to compare the predictive accuracy of competing forecasting systems (or more specifically, tests for the null hypothesis of no difference in the accuracy of two competing forecasts). In [17], too, the predictive accuracy of competing forecasting systems is compared (see also Comment 6 in that paper). An important contribution of the latter work is the introduction of *test functions* to test predictive accuracy *conditionally* on

available information. This is similar to the idea of *stratification*, apparently independently discovered in the meteorological community and to be discussed Section 2.2 (see also references in that section). The methodology in [17] also exploits the finite correlation length of what corresponds to our $\{\phi(k), k \in \mathbb{N}\}$ (see Theorems 1 and 3 of that paper).

In [29], the authors are interested in comparing risk measure estimating procedures. With regard to statistical tests, the paper uses the methodology of [17] in a one step ahead scenario. The work presented in [34,35] considers data-based evaluation of statistical models in a prequential framework (more on this framework below). Again, the authors consider a one-step ahead situation and are thus able to use the martingale property of $S(n)$ in their statistical methodology. An additional point worth noting about [17,35] is that additional dependence of the forecasts on estimated nuisance parameters is considered.

A number of publications [3,4,14,19,20,26,28,31,40] identify positive attributes of probability and ensemble forecasts (such as reliability and resolution) as well as ways to quantify those attributes but do not necessarily suggest a statistical methodology to test or estimate those attributes.

In view of this previous work, the main contribution of the present paper is a rigorous methodology for testing reliability of forecasting systems in particular for larger lead times. Furthermore, the idea of stratification (which originates in the meteorological community) will be rigorously embedded into this methodology, and the connection to test functions will be clarified.

The present contribution also draws on research by Dawid, Vovk, and coworkers who investigated forecast evaluation in a series of papers [11,12,34,35]. As was already mentioned, the authors consider particular forecasting systems with unit lead time only. In addition, however, they formulate what they refer to as the weak and the strong prequential principle which applies to any forecasting system, and we shall discuss these principles briefly. As we have seen, to evaluate a forecasting system in the present framework, we require an archive of verification–forecast pairs, and the desired nominal statistical behaviour of the forecasting system (i.e. what it means for it to be reliable) has to be clear. The prequential principles imply that indeed *only* these two ingredients should matter in forecast evaluation. More specifically, the weak prequential principle says that in terms of actual data, only forecasts that were actually issued and verifications that were actually observed should be used; no other data (for instance, hypothetical forecasts that were in fact never issued) should be taken into account. The strong prequential principle demands that in terms of statistical assumptions, only the nominal forecast behaviour should be used, and no other assumptions should be made about the statistical properties of the verification. If for instance a forecaster claims her forecasts to represent the conditional mean and variance of the verification, it would violate the strong prequential principle to assume, in addition, that the conditional distribution of the verification is normal (with the forecast mean and variance as parameters); in doing so, we would evaluate not only the forecaster's claims but also our own assumptions about the problem, which seems unfair to the forecaster. Naturally, complete adherence to the strong prequential principle in particular can be difficult to achieve.

In Section 2.1, we will fix some notation and introduce a few different classes of forecasting problems, while Section 2.2 discusses the strongly related concepts of stratification and of test functions. In Section 2.3, we present three results that are key to our methodology: a Law of Large Numbers and a Central Limit Theorem regarding the statistic S as

well as a Corollary introducing the eventual (χ^2 or Wald-type) test statistic t^2, which also involves a consistent estimator of the variance of S (which is unknown in general). The Law of Large Numbers and the Central Limit Theorem are stated without precise conditions though. These will be specified in Sections 3.1 and 3.3, adapted to different forecasting situations. Namely, Section 3.1 considers problems with a forecast lead time (or horizon) of a single time step (in a sense to be made precise), largely for the purpose of illustration, while Section 3.3 considers problems with larger lead times. A few numerical examples relating to the situations considered in Sections 3.1 and 3.3 will also be presented (in Sections 3.2 and 3.4, respectively). Section 4 concludes, while proofs to our main theorems can be found in the Supplementary Material. All proofs can be reduced to existing Laws of Large Numbers and Central Limit Theorems for dependent variables (see e.g. [41]), the proof of Proposition 3.2 (covering larger lead times) requiring substantially more work though. The Supplementary Material also contains a brief overview over the software package `franz` [8], which contains python implementations of the methods discussed in this paper.

2. Mathematical methodology

2.1. Notation and basic definitions

We recall that $\{Y(k), k = 1, \dots, n\}$ and $\{f(k), k = 1, \dots, n\}$ denote the verification and the forecasts, respectively. These we model as random variables on some measurable space $(\Omega, \mathcal{A})$, where the $Y(k)$ and $f(k)$ have values in some measurable spaces $(E, \mathcal{A}_E)$ and $(F, \mathcal{A}_F)$, respectively (these will be very simple spaces in the examples discussed later). We assume that for each k the forecast $f(k)$ depends on some information, available to the forecaster at forecast time. We model this by assuming that $\{f(k), k \in \mathbb{N}\}$ is adapted to a filtration $\{\mathcal{F}(k), k \in \mathbb{N}\}$. We remember that k represents the time at which $f(k)$ verifies, so the forecaster will know $\mathcal{F}(k)$ (and issue $f(k)$) at some point in time typically prior to k, for instance at time $k-T$ where T is called the lead time or forecast horizon. An example for $\mathcal{F}(k)$ could be the sigma-algebra generated by the verifications available to the forecaster at the time when she has to issue the forecast $f(k)$ (which in the example above would be $\{Y(l), l \leq k - T\}$). Finally, for each $k \in \mathbb{N}$ and all $A \in \mathcal{A}_E$ we use the shorthand $P_k(A) = \mathbb{P}(Y(k) \in A \mid \mathcal{F}(k))$ for what some authors call the *forecast distribution*.

Definition 2.1: Consider a measurable mapping $\Phi : E \times F \to \mathbb{R}^D$, which we will refer to as the *identification function*. Put $\phi(k) := \Phi(Y(k), f(k))$ for $k \in \mathbb{N}$. We say that the forecasting system is *reliable* (with respect to the identification function Φ) if[1] $\mathbb{E}(\min\{0, \phi(k)\}) > -\infty$ and

$$\mathbb{E}(\phi(k) \mid \mathcal{F}(k)) = 0 \tag{3}$$

for all $k \in \mathbb{N}$. If the forecasting system is reliable, we say that it *standardises* the identification function Φ if $\mathbb{E}(\phi(k)\phi(k)^t \mid \mathcal{F}(k)) = \mathbb{1}$ for all $k \in \mathbb{N}$ (where the superscript t represents the transpose and $\mathbb{1}$ the $D \times D$ unit matrix).

The concept of identification functions has been considered in the literature albeit with a more specific meaning. According to [29], for instance, if we fix a set $\mathcal{P}$ of probability

distributions over $(E, \mathcal{A}_E)$ and a function $r : \mathcal{P} \to \mathbb{R}$, then Φ is an identification function with respect to the *risk measure* r if

$$\int_E \Phi(y,f)\mu(\mathrm{d}y) = 0 \iff r(\mu) = f. \tag{4}$$

We will refer to such identification functions as *explicit* identification functions. Possible risk measures are for instance quantiles or moments. For connections to the *elicitability* or *identifiability* problem, see for instance [18,38].

Assume we are given verifications $\{Y(k), k \in \mathbb{N}\}$ and forecasts $\{f(k), k \in \mathbb{N}\}$ that are reliable with respect to an explicit identification function Φ with respect to a risk measure r. Then, Equation (4) holds with $\mu := P_k$ and $f := f(k)$, and we can conclude that $f(k) = r(P_k)$ for all $k \in \mathbb{N}$; in other words, the forecasting system is reliable if and only if it provides the correct conditional risk measure for all k.

Although strongly related, the identification functions considered in the present paper will not necessarily be explicit identification functions (e.g. in Examples 2.1 and 2.3). There are two reasons for this. Firstly, the reliability condition already implies an implicit relation between $f(k)$ and P_k for all $k \in \mathbb{N}$. Using an explicit identification function merely allows to solve this relation for $f(k)$. In the applications we have in mind, this is often not essential. It will turn out though that, assuming reliability, the identification functions considered in the examples will always allow for representing part of the forecast as a function of other parts of the forecast and the conditional probability.

Secondly, the null hypothesis might involve that the forecasts standardise the identification function. This will simplify estimating the variance of the statistic S and applying the Central Limit Theorem (more specifically the variance estimators in Corollary 2.1). It turns out though that working with explicit identification functions *and* imposing standardisation can result in rather restricted or contrived null hypotheses. Indeed, when working with an explicit identification function, the forecast is fixed to $r(P_k)$ as soon as reliability holds. Assuming in addition that the forecast standardises the identification function, we would have

$$\int_E \Phi(y, r(\mu))\Phi^t(y, r(\mu))\mu(\mathrm{d}y) = \mathbb{1} \tag{5}$$

for any μ which can possibly appear as a value of P_k under the null hypothesis. This can imply strong restrictions on the possible null hypotheses, and there appear to be two ways to avoid this. The first way is to use explicit identification functions with invertible risk measure whenever possible (see Example 2.2 for such a situation). This means that as soon as reliability holds, the forecasts $f(k)$ specify P_k completely. In this case, the left hand side of Equation (5) can be written as a function of the forecast only. Therefore, we can divide the identification function by this quantity and obtain a new explicit identification function which is automatically standardised by any reliable forecast.

The second approach is to drop the requirement that the identification functions are explicit (see Examples 2.1 and 2.3). We should stress however that there are caveats to this second approach. That the forecasts standardise, the identification function is not any longer a consequence of reliability but a nontrivial extension of the null hypothesis. The proposed tests, however, do not necessarily develop a lot of power against violation of standardisation, with the result that the test size on parts of the null hypothesis might be larger

than the power against certain alternatives. This problem will be discussed in the context of Example 2.1. If reliability tests against such alternatives are envisaged, it is advisable to drop the assumption of standardisation. This will require the use of more complicated variance estimators in Corollary 2.1 (the precise statement is Proposition 3.2).

Several examples shall illustrate Definition 2.1 and the subsequent discussion. The tests to be presented later in Section 3 have been implemented in the context of these examples as part of the `franz` software package (see Section 2 in the Supplementary Material for more information about `franz`).

Example 2.1 (Mean and variance forecasts): This example expands on Example 1.1. Again, the verifications are real numbers so that $E = \mathbb{R}$. For each $k \in \mathbb{N}$, the forecast $f(k)$ comprises two numbers $f(k) = (f_1(k), f_2(k))$ where $f_2(k) > 0$. Hence $F = \mathbb{R} \times \mathbb{R}_{>0}$. As an identification function, we take $\Phi(y, f) = \frac{y - f_1}{\sqrt{f_2}}$. The forecasting system is reliable and standardises the identification function whenever $f_1(k)$ and $f_2(k)$ are equal to, respectively, the conditional expectation and the conditional variance of $Y(k)$ given $\mathcal{F}(k)$. The identification function is not explicit since reliability itself will only specify $f_1(k)$ but not $f_2(k)$. It is also clear that a test based on S might exhibit power problems against alternatives that are still reliable (i.e. where f_1 provides the correct conditional expectation), but f_2 is larger than the conditional variance of $Y(k)$.

As already mentioned, testing against such alternatives requires a slightly more complicated test that involves an additional estimation of the variance of S (see Proposition 3.2).

Example 2.2 (Probability forecasts for categorical events): In this example, which expands on Example 1.2, the verifications $\{Y(k), k \in \mathbb{N}\}$ take values in a finite set $E = \{1, \ldots, M\}$, which might, for example, correspond to M mutually exclusive categories of weather. The forecasts $\{f(k), k \in \mathbb{N}\}$ take values in the set of probability vectors over E, that is, the set of M-dimensional vectors $p = (p_1, \ldots, p_M)$ where $p_m \in [0, 1]$ for all $m = 1, \ldots, M$ and $\sum_{m=1}^{M} p_m = 1$. We seek an identification function so that the forecasting system is reliable and standardised if the forecast $f(k)$ represents the conditional probability distribution of $Y(k)$ given $\mathcal{F}(k)$, that is $f_m(k) = \mathbb{P}(Y(k) = m \mid \mathcal{F}(k))$ for all m, k. Motivated by Example 1.2, we could try an identification function Φ with values in $\mathbb{R}^M$ and components $\Phi_m(y, p) = (\delta_{m,y} - p_m)/\sqrt{p_m}$ (reminiscent of Pearson's goodness of fit test). But since $\Phi(y, p) \perp \sqrt{p}$ for any y, p (with $\sqrt{p}$ being understood component-wise), the covariance of $\phi(k)$ would always be rank deficient, having a kernel spanned by $\sqrt{p}$ and rendering standardisation impossible. We remove this problem by projecting onto the subspace orthogonal to $\sqrt{p}$, using the following

Lemma 2.1: *Define the set*

$$S := \left\{ x \in \mathbb{R}^M; x_m \geq 0 \text{ for } m = 1, \ldots, M; \sum_{m=1}^{M} x_m^2 = 1 \right\}.$$

Then there exists an open neighbourhood S_ϵ of S and a smooth mapping $B : S_\epsilon \to \mathbb{R}^{M \times (M-1)}$ such that for every $q \in S_\epsilon$, the columns of $B(q)$ are orthonormal and orthogonal to q.

(The proof, omitted here for brevity, is easy and relies on the well-known fact that applying the Gram–Schmidt procedure to a set of linearly independent vectors is a smooth

operation.) We use the identification function with values in $\mathbb{R}^{M-1}$ (so that $D = M-1$ in this example) and with components $\Phi_d(y,p) = \frac{1}{\sqrt{p_y}} B_{y,d}(\sqrt{p})$ for $d = 1,\ldots,D = M-1$. If $p_y = 0$ we set $\Phi(y,p) = 0$ (note that under the null hypothesis of reliability, the event $p_{Y_k} = 0$ happens with zero probability). Regarding this identification function, we have the following lemma:

Lemma 2.2: *Φ is an explicit identification function with risk measure being the identity on the set of probability vectors over $E = \{1,\ldots,M\}$. That is $\sum_{m=1}^{M} \Phi(m,q)p_m = 0$ for two probability vectors p, q implies $p = q$. Furthermore, for any probability vector p such that $\sum_{m=1}^{M} \Phi(m,p)p_m = 0$, we have*

$$\sum_{m=1}^{M} \Phi_d(m,p)\Phi_{d'}(m,p)p_m = \delta_{d,d'}$$

for $d, d' = 1,\ldots,M-1$.

The second statement implies that a reliable forecast will automatically standardise the identification function. (Again, the proof is easy and omitted for brevity.)

Example 2.3 (Probability forecasts for continuous variables): In this example, $E = \mathbb{R}$ so that the verifications are real numbers, while the forecasts $\{f(k), k \in \mathbb{N}\}$ are continuous cumulative distribution functions (CDF's).

Referring back to Definition 2.1, in the present example, the space F is thus given by the space of all continuous cumulative distribution function over $\mathbb{R}$. We seek identification functions such that the forecasting system is reliable and standardises the identification function if $f(k)$ represents the conditional cumulative distribution function of $Y(k)$ given $\mathcal{F}(k)$, that is $f(k;x) = \mathbb{P}(Y(k) \le x \,|\, \mathcal{F}(k))$ for all k and all $x \in \mathbb{R}$.

For a given $y \in \mathbb{R}$ and a given continuous cumulative distribution function G, we consider an identification function Φ with values in $\mathbb{R}^D$ and components given by

$$\Phi_d(y,G) = \lambda_d(G(y)) \quad \text{for } d = 1,\ldots,D,$$

where λ_d is the Legendre polynomial of degree d on $[0,1]$. These polynomials emerge if the Gram–Schmidt procedure in $L_2([0,1],\mathrm{d}x)$ is applied to the standard monomials $1, x, x^2, \ldots$. If reliability holds, the random variable $R(k) := f(k;Y(k))$, known as the *probability integral transform (PIT)* of $Y(k)$, has a uniform distribution conditionally on $\mathcal{F}(k)$, for all $k \in \mathbb{N}$. It now follows from the properties of the Legendre polynomials that the forecasts are reliable with respect to this class of identification functions and furthermore standardise it. We stress that in general the $\{R(k), k \in \mathbb{N}\}$ are not independent.

The identification function is not explicit since reliability itself will only guarantee that the first D moments of the PIT are those of a uniform distribution but not that the PIT is uniformly distributed. It is possible to construct examples where the first D Legendre polynomials of the PIT have zero expectation even though the PIT is not uniform, and the Legendre polynomials do not have unit covariance matrix. Again, a test based on S might exhibit power problems against such alternatives. As with Example 2.1 one should consider the test of Proposition 3.2 if such alternatives are a possibility.

2.2. Stratification

As was already hinted at in the introduction, a simple performance index like the statistic S (Equation (2)), especially with scalar values, might not develop enough power in order to detect deviations from reliability, since the expected value of $S(n)$ might be zero even if the conditional expectations in Equation (3) are not zero. More specifically, suppose that $\{\phi(k), k = 1, 2, \ldots\}$ are ergodic so that the Law of Large Numbers can be applied to $\frac{1}{n}S(n)$. If the forecasts are reliable, then for $n \to \infty$ we get that $\frac{1}{n}S(n)$ converges almost surely to $\mathbb{E}(\phi(k))$ which is zero as a consequence of Equation (3). The converse conclusion, however, is false: It is possible that $\mathbb{E}(\phi(k)) = 0$ for all k even though reliability (i.e. Equation 3), which is a stronger condition, fails to hold. Examples for such a situation will be considered below and in Section 3.4.

The situation is entirely analogous to rank histograms, a popular tool in the atmospheric sciences to evaluate the reliability of ensemble forecasts. As argued in [22], a single rank histogram might not be sufficient to detect deviations from reliability. To deal with this problem and to build tests that can detect deviations from reliability that do not manifest themselves in single rank histograms, the concept of *stratification* was introduced [see 2, 6, 37]; see also [1] for an in-depth discussion of this technique.

Although developed independently in the meteorological community, the basic idea is analogous to the concept of test functions presented in [17] and will be used here as well in simplified form. Suppose that $Y(k), k = 1, 2, \ldots$ are verifications with $f(k), k = 1, 2, \ldots$ corresponding forecasts that are reliable with respect to an identification function Φ (which we assume for the moment has values in $\mathbb{R}$). We note that the forecasts will still be reliable (i.e. Equation (3) will still hold) with respect to a modified identification function $\tilde{\Phi}$ of the form

$$\tilde{\Phi}(y,f) = \Phi(y,f)\zeta(f), \tag{6}$$

where ζ is any measurable and bounded function which one may call a *test function* (by considering bounded test functions we avoid any integrability issues). Indeed, for any such test function ζ, Equation (3) implies that $\mathbb{E}[\Phi(Y(k),f(k)) \cdot \zeta(f(k))] = 0$ for any k. Hence, modifying the identification function in this way, we can expect to obtain power against alternative hypotheses for which $\mathbb{E}(\Phi(Y(k),f(k))) = 0$ but $\mathbb{E}\left[\Phi(Y(k),f(k)) \cdot \zeta(f(k))\right] \neq 0$.

To illustrate this point, we discuss a special case of Example 1.1. Suppose that forecasts and verifications take real values and are connected through

$$Y(k) = f(k) + r(k), \quad k = 1, 2, \ldots$$

where the forecasts are independent and identically distributed with mean zero, and the $r(k), k = 1, 2, \ldots$ (representing some form of noise) are likewise independent and identically distributed with mean zero, and furthermore independent of the forecasts. We assume that at time k, the forecaster knows $f(k)$ as well as $(Y(l),f(l))$ for all $l < k$ (i.e. these variables generate $\mathcal{F}(k)$). Now indeed Equation (1) is satisfied and the forecasts are reliable if we take $\Phi(y,f) = y - f$ as identification function. We may look at the forecasts $g(k) := \alpha f(k)$ instead, for some $\alpha \neq 1$ and all $k = 1, 2, \ldots$. The filtration $\{\mathcal{F}(k), k = 1, 2, \ldots\}$ remains unchanged but the new forecasts are not reliable. However,

still $\mathbb{E}(\Phi(Y(k), g(k))) = 0$ for all k so that $\frac{1}{n}S(n)$ will converge to zero and the lack of reliability will not be detected. This changes if we introduce the test function $\zeta(g) := \text{sign}(g)$, as now $\mathbb{E}\left[\Phi(Y(k), g(k)) \cdot \zeta(g(k))\right] = (1-\alpha)\,\mathbb{E}|g(k)|$, which is nonzero (unless $\alpha = 1$, in which case the forecast is reliable). Hence $\frac{1}{n}S(n)$ will be significantly different from zero (provided a sufficient amount of data).

The general idea behind stratification is, roughly speaking, to improve the power by using several test functions in parallel. We will further generalise on the ideas outlined above by allowing for identification functions with values in $\mathbb{R}^D$ (as in Definition 2.1) and by allowing for test functions that at time k depend not only on the current forecast $f(k)$ but may be measurable with respect to $\mathcal{F}(k)$, that is, the entire information available at forecast time. More specifically, we consider a randomly weighted sum with values in $\mathbb{R}^{D\times L}$ with the components being of the form

$$S_{d,l}(n) = \sum_{k=1}^{n} \Phi_d(Y(k), f(k)) Z_l(k).$$

Here, the weights $\{Z(k), k \in \mathbb{N}\}$ are adapted to $\{\mathcal{F}(k), k \in \mathbb{N}\}$ and take values in the set of canonical basis vectors $\{e_l, l = 1, \ldots, L\}$ of $\mathbb{R}^L$ for some $L \in \mathbb{N}$. That is, $Z(k)$ is an L-dimensional vector with zero coordinates except for a single unit entry in a random position.

In [17,29], the $\{Z(k), k \in \mathbb{N}\}$ are called *test functions* and permitted to take values in $\mathbb{R}^L$. However, Equation (3) is in fact equivalent to $\mathbb{E}(\phi(k)Z(k)) = 0$ for any weight (or test function) $Z(k)$ with values 0 or 1. So using weights with discrete values does not impose a restriction from that perspective. In the present paper, weights will have discrete values mainly for reasons of interpretability though. The different values assumed by the variable $Z(k)$ should be seen as indicators of different forecasting scenarios, for instance, different synoptic weather patterns in atmospheric contexts or different global economic regimes. A weather forecaster, for example, might be interested in checking whether there is a difference in the performance of the forecasting system under monsoon conditions (or during El Niño) as opposed to the absence of monsoon (or la Niña, respectively). The requirement that $\{Z(k), k \in \mathbb{N}\}$ is adapted to $\{\mathcal{F}(k), k \in \mathbb{N}\}$ is natural as the forecaster would know about these conditions at forecast time.

In case the image of Φ has more than one dimension (i.e. $D > 1$), we multiply each component with the same set of weights. From now on, we will use the shorthand $\psi(k) := (\phi_d(k)Z_l(k))_{d,l}$; note that $\psi(k) \in \mathbb{R}^{D\times L}$. Assuming that the forecasts are reliable with respect to Φ, a straightforward calculation will show that

$$\mathbb{E}(\psi_{d,l}(k) \mid \mathcal{F}(k)) = \mathbb{E}(\phi_d(k)Z_l(k) \mid \mathcal{F}(k)) = \mathbb{E}(\phi_d(k) \mid \mathcal{F}(k))Z_l(k) = 0,$$

while if the identification function is standardised, we get

$$\mathbb{E}(\psi_{d,l}(k)\psi_{d',l'}(k) \mid \mathcal{F}(k)) = \mathbb{E}(\phi_d(k)Z_l(k)\phi_{d'}(k)Z_{l'}(k) \mid \mathcal{F}(k)) = Z_l(k)\delta_{l,l'}\delta_{d,d'}. \qquad (7)$$

We have seen that taking $Z(k)$ to be a function of the forecast $f(k)$ for all $k \in \mathbb{N}$ is a possible stratification as this will render $\{Z(k), k \in \mathbb{N}\}$ adapted to $\{\mathcal{F}(k), k \in \mathbb{N}\}$, and that in this case $\psi(k)$ can be written as a function of the verification and the forecast. Therefore, it can be interpreted as a new identification function (see Equation (6)). This is fine, except that the new identification function will not be standardised, as Equation (7) shows.

2.3. The law of large numbers, the central limit theorem, and a generalised χ^2-test

If the forecasting system is reliable and the identification function is standardised, we would expect the statistic $S(n)/\sqrt{n}$ to be asymptotically normal with a mean zero. More specifically, we would expect the following Law of Large Numbers and Central Limit Theorem to hold. We will formulate these as theorems here albeit without providing the required assumptions yet. Specific sets of assumptions will be stated in Sections 3.1 and 3.3.

Theorem 2.1 (Law of Large Numbers): *If the forecasting system is reliable and under additional conditions stated in Sections 3.1 and 3.3, we have*

$$\frac{1}{n}S(n) = \frac{1}{n}\sum_{k=1}^{n}\psi(k) \to 0 \quad \text{for } n \to \infty.$$

The convergence takes place in mean square and with probability one.

Theorem 2.2 (Central Limit Theorem): *If the forecasting system is reliable and under additional conditions stated in Sections 3.1 and 3.3, the quantity*

$$\frac{1}{\sqrt{n}}S(n) = \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\psi(k)$$

converges in distribution (for $n \to \infty$) to a normal distribution with mean zero and covariance v^2.

The Law of Large Numbers provides a justification for using the statistic S as a means of assessing reliability and thus of the entire methodology. According to the Central Limit Theorem, we can refer $\frac{1}{\sqrt{n}}S(n)$ to the normal distribution in order to assess whether the deviations from zero are unduly large. A final problem presents itself with applying this in practice, namely, that the variance v referred to in the Central Limit Theorem is in general not known (even if the forecasting system is standardised, we will have $v \neq \mathbb{1}$ in general). The following corollary of Theorem 2.2 is, therefore, relevant.

Corollary 2.1 (Generalised χ^2-test or Wald-type test): *Assume the Central Limit Theorem (Theorem 2.2) holds and v is positive definite*[3]*. If for each k, there are estimators $v(k)$ (measurable with respect to $\mathcal{F}(k)$) so that $v(k) \to v$ when $k \to \infty$, where v is as in the Central Limit Theorem, then*

$$t^2 := \frac{1}{n}S(n)^t v(n)^{-1} S(n) \tag{8}$$

has asymptotically a χ^2-distribution with $D \cdot L$ degrees of freedom.

(For some $w \in \mathbb{R}^{(D\times L)\times(D\times L)}$, the inverse w^{-1} is an element from the same space such that $\sum_{k,l}(w^{-1})_{i,j,k,l}w_{k,l,m,n} = \delta_{i,m}\delta_{j,n}$ holds.) In Sections 3.1 and 3.3, we will discuss the precise assumptions to these theorems in the context of Examples 2.1–2.3. In particular, we will illustrate the usage of the Generalised χ^2-test (Corollary 2.1) and provide explicit formulae for estimators of the variance v (see Propositions 3.1, 3.2). All proofs are deferred to Section 1 in the Supplementary Material.

3. Application: forecasts with specified lead time

3.1. Forecasts with unit lead time

We start with the observation that if $Y(k)$ were $\mathcal{F}(k)$ -measurable, then the verification would be a function of the information available at forecast time and hence not any more uncertain. Although finding this function might still be difficult in practice, here we are interested in 'truly random problems' where $Y(k)$ is not $\mathcal{F}(k)$-measurable for any k. In many practical situations, it is true though that

$$Y(k) \text{ is } \mathcal{F}(k+T)-\text{measurable for some } T > 0 \text{ and all } k. \tag{9}$$

That is, when forming the forecast $f(k+T)$, the forecaster knows $Y(k)$. An example of such a situation is if verification and forecast information are obtained from the same observational network, but the forecasts have a forecast horizon or *lead time* T.

In this section, we assume $T = 1$. The next section will deal with larger lead times, that is $T > 1$. As mentioned in the introduction, the papers [34,35] deal exclusively with the situation of lead time $T = 1$. (In fact, the authors make the even stronger assumption that $\mathcal{F}(k)$ is the sigma-algebra generated by $Y(1), \ldots, Y(k-1)$, but this does not afford much simplification of the analysis.) Therefore, the present section mainly serves the purpose of illustration, while originality is limited to considering a wider range of identification functions than those papers.

We assume that the forecasting system is reliable and standardises the identification function. Condition (9) with $T = 1$ then leads to a very strong decorrelation property of the $\{\psi(n), n \in \mathbb{N}\}$. Indeed, from Condition (9), we obtain that in fact $Y(k)$ is $\mathcal{F}(n)$-measurable for any $k \leq n-1$ and since the same is true for $f(k)$ and $Z(k)$, we can conclude that the $\psi(k)$ are $\mathcal{F}(n)$-measurable for any $k \leq n-1$. In other words, when issuing $f(n)$, the forecaster knows $\psi(k)$ for all $k = 1, \ldots, n-1$. The law of the iterated expectation in conjunction with reliability implies for all k:

$$\mathbb{E}(\psi(k) \mid \psi(n), n = 1, \ldots, k-1) = \mathbb{E}\big[\mathbb{E}(\psi(k) \mid \mathcal{F}(k)) \mid \psi(n), n = 1, \ldots, k-1\big] = 0.$$

This means that $\{S(n), n \in \mathbb{N}\}$ is a martingale and $\{\psi(n), n \in \mathbb{N}\}$ a process of martingale differences or a *fair* process [see for instance, the books [9,15], which contain introductory chapters on the subject]. In particular, fair processes satisfy Laws of Large Numbers and Central Limit Theorems, under appropriate additional conditions:

Proposition 3.1: *We assume the following conditions*

(i) *The forecasting system is reliable.*
(ii) *The forecasting system standardises the identification function.*
(iii) *Condition* (9) *is satisfied with* $T = 1$.

Then the Law of Large Numbers (*Theorem* 2.1) *holds. If in addition we have*

(iv) $\{Z(k), k \in \mathbb{N}\}$ *is an ergodic process with* $q_l := \mathbb{E}(Z_l(k)) > 0$ *for all* $l = 1, \ldots, L$.

(v) *The following Lindeberg condition holds:*

$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}(\psi^2(k)\mathbb{1}_{\psi^2(k)\geq\epsilon n}\,|\,\mathcal{F}(k)) \to 0 \text{ in probability}$$

for all $\epsilon > 0$.

Then the Central Limit Theorem (*Theorem* 2.2) *holds with a covariance* v *having components*

$$v_{d,l,d',l'} = q_l\delta_{d,d'}\delta_{l,l'}. \tag{10}$$

Further, still under the same conditions, the Generalised χ^2*-test* (*Corollary* 2.1) *is in force, and an estimator for* v *is provided by replacing* q_l *in Equation* (10) *with the following estimator:*

$$\hat{q}_l = \frac{1}{n}\sum_{k=1}^{n} Z_l(k). \tag{11}$$

The proof, which can be found in Section 1 of the Supplementary Material, is a direct application of the Martingale Law of Large Numbers and Central Limit theorems. See also [34] for a similar statement under a slightly different set of assumptions. Note that $\hat{q}_l$ is simply the observed relative frequency of the event $Z_l(k) = 1$ or 'being in stratum l'. The Lindeberg condition is satisfied if Φ is bounded, which is the case in Examples 2.2 and 2.3 (thanks to Lemma 2.1 in the former case). Alternatively, the Lindeberg condition is satisfied if $\{\psi(k), k \in \mathbb{N}\}$ is square integrable and stationary. To see this, let $\Lambda(\epsilon, n)$ for fixed $n \in \mathbb{N}$ and $\epsilon > 0$ be the random variable on the left hand side of the Lindeberg condition. Due to stationarity, we have $\mathbb{E}(\Lambda(\epsilon, n)) = \mathbb{E}(\psi^2(1)\mathbb{1}_{\{\psi^2(1)\geq\epsilon n\}})$ and this converges to zero for $n \to \infty$ since $\psi(1)$ is square integrable. Hence $\Lambda(\epsilon, n) \to 0$ for any $\epsilon > 0$ even in mean square sense and thus also in probability.

It is also worth noting that in Example 2.3 and in the context of the current section (i.e. if Condition (9) is satisfied with $T = 1$), it can be shown that the PITs $\{R(k), k \in \mathbb{N}\}$ are even independent. This is a classical result due to [33] in the case where $\mathcal{F}(k) = \sigma(Y(1), \ldots, Y(k-1))$ for all k, but is easily seen to remain true in the current situation. As the distribution of $\{R(k), k \in \mathbb{N}\}$ is furthermore uniform over $[0, 1]$ in Example 2.3, applying a standard test for the uniformity of the distribution of the $\{R(k)\}$ (such as a Kolmogorov–Smirnov test) constitutes a test for reliability. A test of this form, however, would not involve any stratification and therefore merely assess whether the unconditional (rather than the conditional) distribution of the $\{R(k)\}$ is uniform.

3.2. Numerical examples with unit lead time

To illustrate the methodology outlined in this paper, we will discuss a few numerical experiments, relating to the Examples 2.1–2.3. In the present subsection, we will mainly introduce the systems on which the methodology is to be tested and perform a few experiments for unit lead time; more experiments for larger lead times that will further illustrate the methodology will be discussed in Section 3.4.

The first forecasting system concerns verifications from a one-dimensional autoregressive process (AR-process) of order one, given by

$$Y(n+1) = aY(n) + R(n), \quad n \in \mathbb{N},$$

with $a = 0.5$ and $\{R(n), n \in \mathbb{N}\}$ standard normal and independent. Further, $Y(0)$ is normal with mean zero and variance $\frac{1}{1-a^2}$. We consider conditional mean and variance forecasts for these verifications as discussed in Example 2.1. More specifically, $f_1(k) = aY(k-1)$ and $f_2(k) = 1$ for all $k \in \mathbb{N}$. The forecasts were divided into two strata along the conditional mean $\{f_1(k), k \in \mathbb{N}\}$, using a threshold of zero. This means we set

$$Z_l(k) = \begin{cases} \delta_{l,1} & \text{if } f_1(k) < c \\ \delta_{l,2} & \text{if } f_1(k) \geq c \end{cases} \tag{12}$$

for $c = 0$. In all experiments, $n = 600$ time instances were considered.

The second forecasting system uses synthetic data from a caricature model for monthly precipitation over a region. In this model, the time index k represents months, and the verification $Y(k)$ represents precipitation averages for the corresponding month. The precipitation depends on an underlying random variable $X(k)$ representing the 'climate' of the region which is modelled as a Markov process with three states $\{1, 2, 3\}$ representing 'dry', 'normal', and 'wet' climate, respectively. The distribution of $Y(k)$ given $X(k)$ is a gamma distribution with a density of the form

$$\gamma(y, X(k)) = \frac{1}{2\theta(X(k))^2} y \exp\left(-\frac{y}{\theta(X(k))}\right)$$

with $\theta(x) = 2^{x-1}$. The Markov process $X(k)$ representing the climate has a nonhomogenous (but periodic in time) transition matrix $P(k) = P_{i,j}(k)$ (representing the probabilities of transition from i to j) given by

$$P(k) = \begin{pmatrix} 3/5 & 2/5 & 0 \\ c(k) \cdot 4/5 & 1/5 & (1 - c(k)) \cdot 4/5 \\ 0 & 4/5 & 1/5 \end{pmatrix}.$$

where i110 $c(k) = \frac{1}{2} + \frac{1}{2}\sin(\omega k)$, $\omega = \frac{2\pi}{12}$. Basically, if the system is in the normal state, it either stays put or jumps to one of the other states with a probability of 4/5. Whether a jump ends up in the wet or the dry state depends on probabilities alternating with a seasonal cycle $c(k)$ in opposite directions (i.e. $p_{21}(k) + p_{23}(k) = \text{const.}$). If the system is in the wet state, it either stays put or jumps to the normal state with a probability of 4/5. If the system is in the dry state, it either stays put or jumps to the normal state with a probability of 2/5. A typical time series from this model is shown in Figure 1, along with a one-year running average. As this system constitutes a Hidden Markov model, using the theory of filtering, forecasts for this system can be constructed in the form of cumulative distribution functions, representing the probability distribution of $Y(k)$ given previous observations $Y(k), \ldots, Y(k-T)$ (with $T = 1$ in the present section). By construction, these cumulative distribution functions are reliable in the sense of Example 2.3. Using elementary probability calculus, we can likewise construct mean and variance forecasts and probability forecasts for categorical events. We used the three categories $Y(k) < 2, 2 \leq Y(k) < 4$, and $4 \leq Y(k)$.

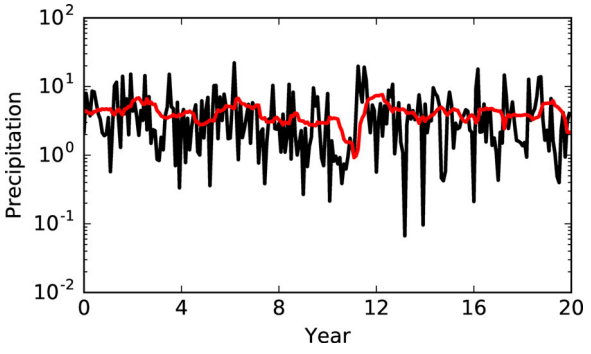

Figure 1. A typical time series from the precipitation model. The ordinate shows amounts of precipitation (note the logarithmic scale), while the abscissa shows time in years. The red line (light grey in print) shows the running average over the previous 12 months.

Again by construction, these forecasts are reliable in the sense of Examples 2.1 and 2.2, respectively. Further, they standardise the employed identification functions.

The forecasts were divided into two strata along the mean forecast (i.e. the conditional expectation under reliability) as for the AR process; see Equation (12) but with $c = 3.98$. This gives roughly equal population of the strata. In all experiments, $n = 600$ time instances were considered.

We note that the conditions of Proposition 3.1 are met both for the AR model as well as the precipitation model, except that strictly speaking $\{Z(k), k \in \mathbb{N}\}$ are not ergodic for the precipitation model since the underlying Markov process is not homogenous. The ergodicity is only needed though to prove that the observed strata populations converge to their expected values, which seems plausible also for the present case as the Markov process has periodic dynamics.

To confirm that the methodology performs as would be expected from Proposition 3.1 (also in case of the precipitation model), we repeated each experiment 1000 times independently, computing the t^2-statistic from the Generalised χ^2-test (Equation 8 in Corollary 2.1) and eventually the p-value every time (using a χ^2-distribution with the appropriate number of dof's). According to our theory, the realisations of the p-value should follow a uniform distribution, at least for data from the AR model. This was indeed confirmed for all four examples, using a Kolmogorov–Smirnov test.

3.3. Forecasts with larger lead times

In this section, we assume that Equation (9) holds albeit with some $T > 1$, essentially corresponding to situations with larger forecast lead time. We obtain that $Y(k)$ is $\mathcal{F}(n)$-measurable for any $k \leq n - T$ and since the same is true for $f(k)$, we can conclude that $\psi(k)$ is $\mathcal{F}(n)$-measurable for any $k \leq n - T$. The law of the iterated expectation in conjunction with reliability then implies

$$\mathbb{E}(\psi(k) \mid \psi(n), n = 1, \ldots, k - T) = \mathbb{E}\left(\mathbb{E}(\psi(k)|\mathcal{F}(k))|\psi(n), n = 1, \ldots, k - T\right) = 0 \tag{13}$$

for all k. In contrast to the previous sections though, the time series $\{\phi(k), k \in \mathbb{N}\}$ is no longer a fair process. Equation (13) does not yield information about $\mathbb{E}(\psi(k) \mid \psi_{1:k-l})$ for

$l = 1, \ldots, L-1$, meaning that we have less control over the correlation structure of the time series $\{\phi(k), k \in \mathbb{N}\}$; the consequence is that the conditions of Proposition (3.1) are no longer sufficient to guarantee the Central Limit Theorem, and compensating assumptions have to be made to care for this lack of information.

Proposition 3.2: *We assume the following conditions:*

(i) *The forecasting system is reliable.*
(ii) $\{\psi(k), k \in \mathbb{N}\}$ *are square integrable and ergodic.*

Then the Law of Large Numbers (*Theorem* 2.1) *holds. If in addition we have*

(iii) *Condition (9) holds for some L.*

Then the Central Limit Theorem (*Theorem* 2.2) *holds. If furthermore*

(iv) *The covariance v is positive definite.*

Then the following estimator for the covariance (with $d, d' = 1, \ldots, D$ *and* $l, l' = 1, \ldots, L$*)*

$$\hat{v}_{d,l,d',l'} = \frac{1}{n}\sum_{k=1}^{n}\left\{\psi_{d,l}(k)\psi_{d',l'}(k) + \psi(k)_{d,l}\left(\sum_{k'=1}^{T-1}\psi_{d',l'}(k+k')\right) + \psi(k)_{d',l'}\left(\sum_{k'=1}^{T-1}\psi_{d,l}(k+k')\right)\right\} \tag{14}$$

satisfies the requirements of Corollary 2.1 *and therefore, the corresponding Generalised* χ^2*-test from Corollary* 2.1 *is valid.*

Remark 3.1: The conditions of Proposition 3.2 are stronger than that of Proposition 3.1 only in that now the ergodicity of $\{\psi(k), k \in \mathbb{N}\}$ is assumed (already in the Law of Large Numbers), and that the covariance v is required to be non-degenerate (see also Remark 3.6).

Remark 3.2: Note that the third term on the right is just the second term but with (d, l) and (d', l') interchanged.

Remark 3.3: We stress that for lead times larger than one, the variance v in the Central Limit Theorem will need to be estimated even if the forecasts standardise the identification function (this will also be confirmed in the numerical examples). The only benefit of standardisation is then that the very first term on the right hand side of Equation (14), which estimates the covariance of $\psi(k)$, can be replaced by the estimator from Proposition 3.1, Equation (11) for the case of unit lead time. In particular, this term will be diagonal, that is, nonzero only if $d = d'$ and $l = l'$ so only these entries have to be estimated.

Remark 3.4: As has been mentioned several times, Proposition 3.2 has to be used even for lead time $T = 1$ if standardisation is not part of the null hypothesis.

Remark 3.5: The ergodicity of $\{\psi(k), k \in \mathbb{N}\}$ cannot usually be inferred in a real-world example, and is in fact wrong in the context of the synthetic precipitation model since this model is not time homogenous. As discussed, this implies that by invoking the Central Limit Theorem, we are assessing the consistency between the forecaster's claims and the data under the *additional* assumption that $\{\psi(k), k \in \mathbb{N}\}$ is ergodic.

Remark 3.6: It would be desirable to have primitive conditions guaranteeing that the covariance v is non-degenerate. Unfortunately, this does not seem to be a universal property under the null hypothesis of reliability but rather is dependent on the precise correlation structure of the time series $\{\psi(k), k \in \mathbb{N}\}$. Nor is it obvious that non-degeneracy of v can be guaranteed under reasonably general conditions by suitable design of the identification function. A step in that direction is identification functions that are standardised by the forecasts. This will at least guarantee the covariance of $\psi(k)$ to be invertible, and this often appears to be the dominant contribution to v.

3.4. Numerical examples with larger lead times

Forecasts for both the precipitation model as well as the AR model can be constructed for higher lead times as well. As before, mean and variance forecasts (for both systems), probability forecasts for categorical events, and cumulative distribution forecasts (for the precipitation model) were generated that are reliable in the sense of the corresponding Examples 2.1–2.3; the forecasting systems furthermore standardise the employed identification functions. Assuming also ergodicity, Proposition 3.2 is in force, and we estimate v using estimators $v(n)$ from Equation (14). All experiments were conducted with $n = 600$ time instances and for various lead times; stratification (whenever present) was performed in the same way as for unit lead time.

As in the previous section, we repeated the experiments 1000 times, each time computing the test statistic t^2 for these forecasting systems (including estimators for the variance). Despite the fact that strictly speaking, the precipitation model is not ergodic and hence Proposition 3.2 is not applicable, we nonetheless find that realisations of the t^2-statistic from these experiments exhibit a distribution that does not significantly deviate from the χ^2 -distribution, or in other words, the p-values follow a uniform distribution as is confirmed through Kolmogorov–Smirnov tests.

For the remainder of this section, we will discuss a few more experiments to illustrate the methodology and the information it provides about forecasting systems. Firstly, we present a typical estimate of the covariance v. It was obtained for cumulative distribution forecasts for the precipitation model with lead time 4 and Legendre polynomials of order up to 6, but no stratification. The covariance is thus a 6×6 matrix, and as estimate (for the correlation matrix) we obtain

$$\frac{\hat{v}_{i,j}}{\sqrt{\hat{v}_{i,i}\hat{v}_{j,j}}} = \begin{pmatrix} 1.00 & -0.05 & -0.09 & -0.10 & 0.12 & -0.05 \\ -0.05 & 1.00 & -0.06 & -0.02 & 0.18 & 0.08 \\ -0.09 & -0.06 & 1.00 & 0.07 & 0.04 & -0.04 \\ -0.10 & -0.02 & 0.07 & 1.00 & -0.08 & -0.12 \\ 0.12 & 0.18 & 0.04 & -0.08 & 1.00 & 0.09 \\ -0.05 & 0.08 & -0.04 & -0.12 & 0.09 & 1.00 \end{pmatrix}, \tag{15}$$

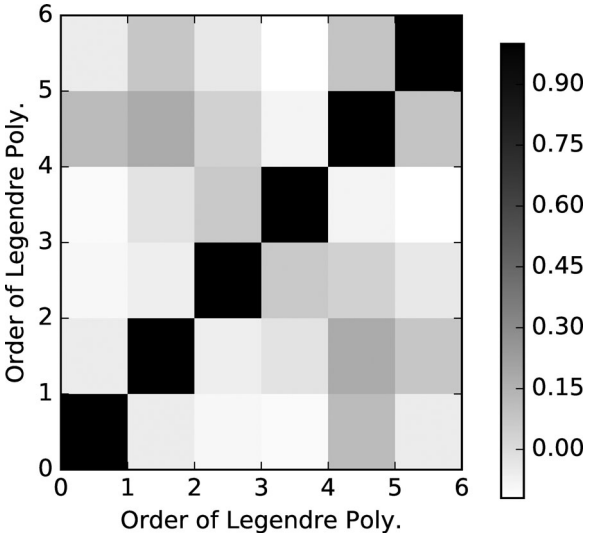

Figure 2. A typical estimate of the covariance v (in fact the correlation matrix, see Equation (15)). It was obtained for cumulative distribution forecasts for the precipitation model with lead time 4 and Legendre polynomials of order up to 6, but no stratification.

and a grey-scale plot is shown in Figure 2. For unit lead time, this would be the unit matrix, but here we seem to have genuine correlations off the diagonal (although we have not conducted a true significance test).

The next experiment shows that these correlations cannot be ignored in the test. The experiment considers categorical forecasts for the precipitation model with a lead time of 4 months. The t^2-statistic was computed assuming no correlations in the time series $\{\psi(n), n \in \mathbb{N}\}$ (or equivalently assuming $v = \mathbb{1}$). This experiment was repeated 1000 times independently; a histogram of the p-values is shown in Figure 3. It is evident that the p-values tend to be too small (a Kolmogorov–Smirnov test for uniformity produces a p-value of $5.48 \cdot 10^{-6}$). Hence, ignoring the correlations, our test would reject a reliable forecasting system too often. Incorporating the correct correlations in the t^2 -statistic, however, produces the correct (i.e. uniform) distribution of the p -values (data not shown), as for the experiments discussed at the beginning of this section. Figure 4, left column, shows results from experiments using cumulative distribution forecasts, moment forecasts, and categorical forecasts for the precipitation model with a lead time of 4 months. The forecasts have been stratified into two categories. The experiments are identical to the corresponding experiments discussed at the beginning of this section, except that now the t^2-statistic was computed assuming no correlations in the time series $\{\psi(n), n \in \mathbb{N}\}$. Top and middle panels of Figure 4, left column, show results for cumulative distribution forecasts and moment forecasts, respectively, while the bottom panel refers again to categorical probability forecasts. The conclusions are the same as before; although there is less evidence that the p-values are not uniform (a Kolmogorov–Smirnov test for uniformity produces larger p -values, see Figure), it would still seem reckless to ignore the correlations. The next experiment demonstrates the value of stratification. In essence, it shows that an unreliable forecasting system might pass an unstratified reliability test, while deviations from reliability become apparent only under stratification. We focus on mean-variance

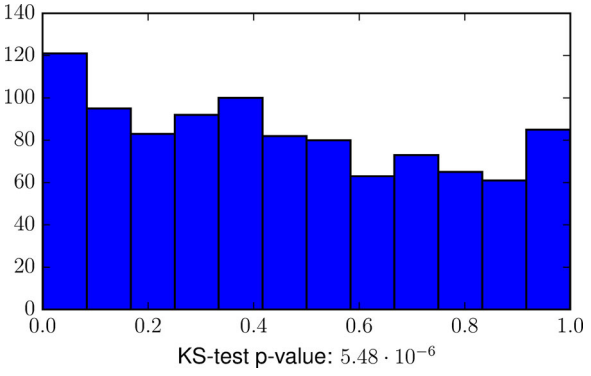

Figure 3. Categorical forecasts for the precipitation model with lead time 4 were tested for reliability (the forecasts are reliable by construction). The t^2-statistic was computed assuming no correlations. Histogram of pvals for 1000 repetitions of this experiment are shown. It is evident that the *p*-values tend to be too small (a KS-test gives a *p*-value of $5.48 \cdot 10^{-6}$). Hence, ignoring the correlations, our test would reject a reliable forecasting system too often.

forecasts for the AR process. Let $\{(f_1(k), f_2(k)), k \in \mathbb{N}\}$ be reliable conditional means and variance forecasts for this problem. We introduce another mean-variance forecasting system $\{(g_1(k), g_2(k)), k \in \mathbb{N}\}$, constructed in such a way that the mean $g_1(k)$ exhibits too little variability, while the variance $g_2(k)$ correctly accounts for the error in the mean $g_1(k)$. More specifically, we set

$$g_1(k) := \alpha f_1(k), \quad g_2(k) := \mathbb{E}((Y(k) - g_1(k))^2) = \frac{1 + (\alpha^2 - 2\alpha)a^4}{1 - a^2}, \tag{16}$$

with $\alpha = 0.4$. The definition of $g_2(k)$ ensures that it reflects the overall (unconditional) expected error of $g_1(k)$. The explicit representation of $g_2(k)$ in Equation (16) follows from a simple calculation. Note that $g_2(k)$ like $f_2(k)$ is in fact independent of k but larger than the latter. We stress that $(g_1(k), g_2(k))$ is *not* reliable since $\mathbb{E}(Y(k) \mid g_1(k)) = \mathbb{E}(Y(k) \mid f_1(k)) = f_1(k) \neq g_1(k)$.

Histograms for the *p*-values from 1000 independent repetitions of the experiment are shown in Figure 4, right column. The top panel refers to tests for reliability without stratifying the forecasts. Evidently, the test is unable to detect lack of reliability as there is no evidence for the *p*-values deviating from a uniform distribution; in other words, the test develops no significant power. Adding stratification though reveals that the forecast is not reliable, as is shown in the middle panel of Figure 4, right column. Now the test develops very significant power. The bottom panel repeats the (stratified) experiment but for the reliable forecasting system, confirming again that the test has the correct size.

As a final application, we consider temperature measurements from a weather station near Bremen, Germany, taken daily between 1 January 2010 and 31 December 2011 at 12 noon (resulting in 730 values). The measurements were converted to anomalies by subtracting a *climate normal* of the form

$$c(k) = c_1 + c_2 \cos(\omega k) + c_3 \sin(\omega k) \quad \text{where } \omega = \frac{2\pi}{365.2425}.$$

The coefficients c_1, c_2, c_3 were found by a least squares fit onto temperature data from the same station but from previous years. The anomalies $\{a(k), k \in \mathbb{N}\}$ were converted to three

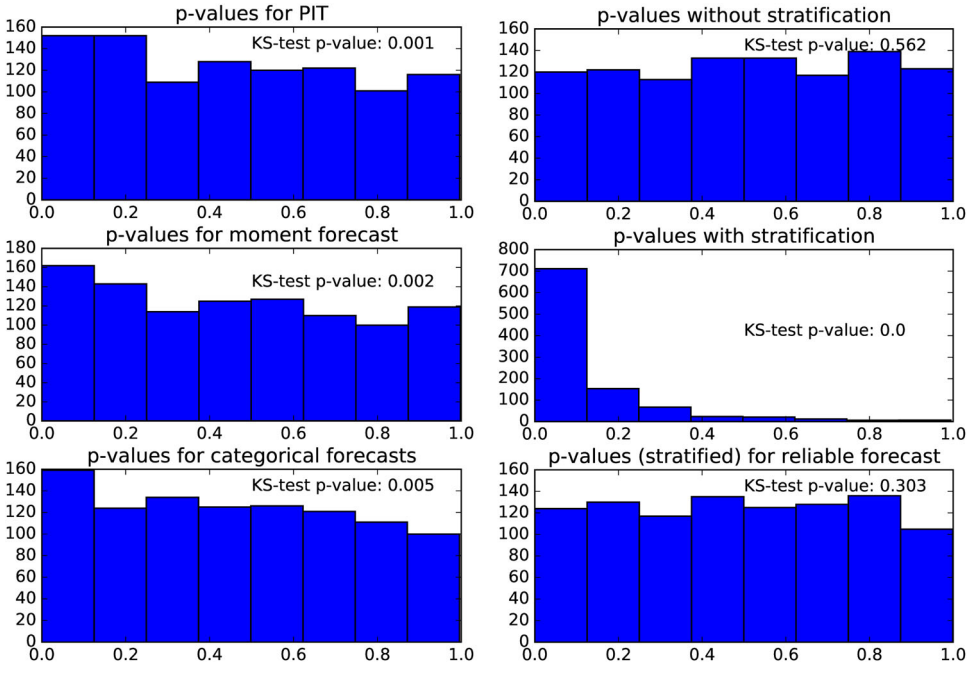

Figure 4. *Left column:* The experiment is similar to Figure 3, except that here the forecasts have been stratified into two categories. Top and middle panels show results for cumulative distribution forecasts and moment forecasts, respectively, while the bottom panel refers again to categorical probability forecasts. Although there is less evidence that the *p*-values are not uniform ignoring the correlations still seems to result in too small *p*-values. *Right column:* A mean-variance forecasting system for the autoregressive process was tested for reliability. The forecasting system was constructed so that the mean exhibits too little variability, while the variance correctly accounts for the error in the mean. Histograms for the *p*-values from 1000 independent repetitions of the experiment are shown. Without stratifying the forecasts, the test is unable to detect lack of reliability (*p*-values exhibit uniform distribution, top panel). Adding stratification though reveals that the forecast is not reliable (middle panel). The bottom panel repeats the (stratified) experiment but for the reliable forecasting system.

categories: $a(k) < -1.867$, $-1.867 \leq a(k) < 1.654$ and $1.654 \leq a(k)$. Each category has a climatological probability of about 1/3.

Forecasts were obtained from the medium-range ensemble prediction system of the European Centre for Medium-Range Weather Forecasts (ECMWF)[4]. The system produces daily ensemble forecasts for the global atmosphere and comprises 50 ensemble members. We consider the forecasts with a lead time of 120 h, which corresponds to $T = 5$. The ensembles are then converted to anomaly ensembles (by subtracting the climate normal) and eventually to probabilities for the three categories as follows: If we let $N_m(k)$ be the number of ensemble members falling into category m at time k, then the forecast probability $f_m(k)$ for category m at time k is found by a (slightly regularised) frequency estimator

$$f_m(k) := \frac{N_m(k) + 1/3}{51}, \quad m = 1, 2, 3.$$

Note that $f_m(k) \geq 1/153$ in any case.

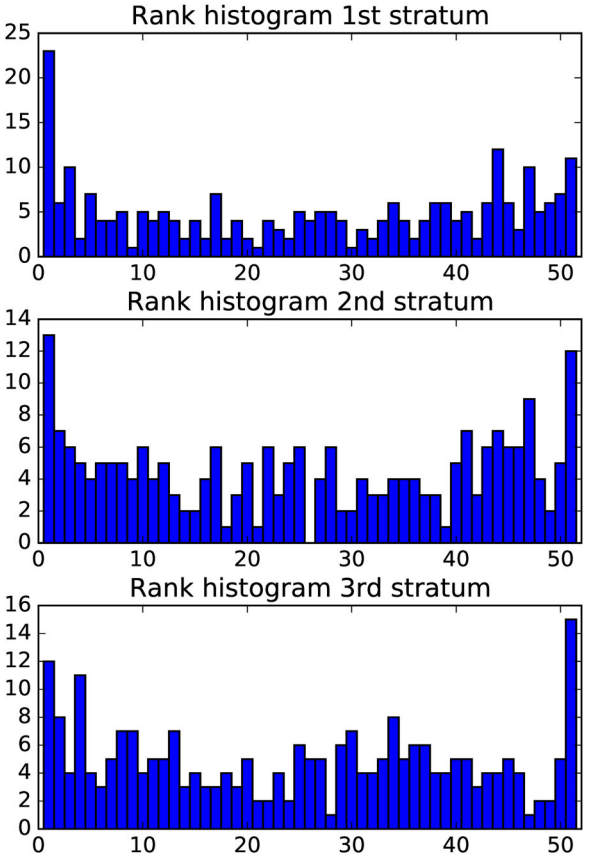

Figure 5. Stratified rank histograms for ECMWF ensemble forecasts for a location near Bremen, lead time 120 h (see][37). Even without a quantitative test, it is evident that these histograms are not 'flat', indicating a lack of reliability in this forecast. Categorical probability forecasts derived from these ensembles though have less resolution and might still be reliable.

For this experiment, we stratify along a variable internal to the forecasting system that indicates 'cold', 'medium' and 'warm' synoptic situations. All three strata contain about the same amount of instances. Testing for reliability with this stratification gives a p-value of 0.584, while testing without stratification gives 0.171. This indicates that there are potential problems with reliability (from the unstratified test) but maybe too little data to see this also in the stratified test.

Stratified rank histograms for the entire ensemble forecasting system are shown in Figure 5. For an introduction to this methodology, see [37]. For a reliable ensemble forecasting system, these histograms should exhibit a uniform distribution, although again temporal correlations need taking into account, see [5]. It is evident however (even without a quantitative test) that these histograms are not 'flat', indicating a lack of reliability in this forecast. This lack of reliability need not necessarily cause the categorical probability forecasts to be unreliable, so there is no inconsistency with the fact that we do not see strong evidence for lack of reliability in this experiment.

4. Conclusion and future work

In this work, we revisited the problem of statistical evaluation of forecasting systems. Central to the presented framework is the concept of nominal forecast behaviour or reliability, which is a statement regarding the supposed statistical properties of the forecasts. Such statements can be interpreted as a hypothesis, and a framework for testing such hypotheses was presented and linked with reliability.

As was noted previously by several authors, the temporal dependence structure of the verification–forecast pairs is unknown in general, but it bears strongly on the distribution of the statistic S (whereby we quantify deviation from reliability), for instance the variance. As we have demonstrated though, the nominal behaviour of the forecasts (i.e. reliability) in fact imposes strong constraints on the dependence structure. More specifically, we showed that with an appropriately chosen identification function Φ, the statistic S becomes a sum over terms with strong decorrelation properties. In forecasting problems with unit lead time, the statistic S even forms a Martingale, a fact that has been used in previous work to identify the asymptotic distribution of S (and related test statistics) using Martingale Central Limit theorems. In the assessment of predictive cumulative distribution functions with unit lead time, the probability integral transform can be used, in which case the statistic S is even a sum over independent and identically distributed random variables.

For larger lead times, the statistic S ceases to be a Martingale but still exhibits strong decorrelation properties. These can be used to show rigorously that the statistic S still obeys the Law of Large Numbers and the Central Limit theorem, and also to find estimators for the correct variance, in general under the additional assumption that the verification–forecast pairs form an ergodic process.

Numerical examples were conducted to demonstrate the validity of the theory and to illustrate its applicability. Synthetic data from a toy model representing a climate region was used to test whether the theory is working. In that experiment, the forecasts are known to be reliable, and the tests indeed behave as predicted under the null hypothesis. We also demonstrate that the non-vanishing correlations have to be taken into account when estimating the variance of the statistic S (as discussed in Section 3.3). Using the classical estimator instead (which ignores correlations) will lead to incorrect behaviour of the test, with too large rejection rates in the discussed example.

Further, we discussed an example demonstrating how stratification of forecasts leads to more powerful tests. A (by construction) unreliable forecasting system was considered that was nonetheless able to pass a test of reliability without stratification. Lack of reliability was detected however if simple stratification was added.

Finally, the difficulty of finding explicit identification functions which are standardised was noted; we were able to find such identification functions only if the risk measure was invertible and thus effectively determined the forecast distribution completely. It would be interesting to know if invertibility of the risk measure is indeed necessary in this situation.

Notes

1. This minimal integrability condition ensures that subsequent expectation values are well defined.
2. Note that for random variables X with values in $\mathbb{R}^{D\times L}$, the covariance ν will be an element of $\mathbb{R}^{(D\times L)\times(D\times L)}$ with components $\nu_{d,l,d',l'} = \mathbb{E}(X_{d,l}X_{d',l'})$.

3. Throughout this paper, *positive definite* means symmetric with positive eigenvalues, in particular implying invertibility.
4. We are grateful to ECMWF and Zied Ben Bouallègue for kindly providing the data.

Acknowledgments

Fruitful discussions with Tobias Kuna, Zied Ben Bouallègue, Stéphane Vannitsem, Valerio Lucarini, Stefan Siegert, Chris Ferro, David Stephenson, and Tobias Fissler are gratefully acknowledged. A number of insightful comments by A. Philip Dawid as well as two anonymous referees and an Associate Editor significantly improved the paper. Forecast and verification data for an example were kindfully provided by the European Centre for Medium-Range Weather Forecasting.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

J. Bröcker http://orcid.org/0000-0002-0864-6530

References

[1] J. Bellier, I. Zin, and G. Bontron, *Sample stratification in verification of ensemble forecasts of continuous scalar variables: potential benefits and pitfalls*, Monthly Weather Rev. 145 (2017), pp. 3529–3544.
[2] J. Bröcker, *On reliability analysis of multi-categorical forecasts*, Nonlinear. Process. Geophys. 15 (2008), pp. 661–673. http://www.nonlin-processes-geophys.net/15/661/2008/.
[3] J. Bröcker, *Reliability, sufficiency, and the decomposition of proper scores*, Q. J. R. Meteorol. Soc. 135 (2009), pp. 1512–1519.
[4] J. Bröcker, *Evaluating raw ensembles with the continuous ranked probability score*, Q. J. R. Meteorol. Soc. 138 (2012), pp. 1611–1617. http://dx.doi.org/10.1002/qj.1891.
[5] J. Bröcker, *Assessing the reliability of ensemble forecasting systems under serial dependence*, Q. J. R. Meteorol. Soc. 144 (2018), pp. 2666–2675. https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3379.
[6] J. Bröcker and Z. Ben Bouallègue, *Stratified rank histograms for ensemble forecast verification under serial dependence*, Q. J. R. Meteorol. Soc. 146 (2020), pp. 1976–1990. https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3778.
[7] J. Bröcker, *Probability forecasts*, in Jolliffe and Stephenson [24], chap. 8, pp. 119–139.
[8] J. Bröcker, `franz`, *a python library for statistical assessment of forecasts (release 1.0)*, GitHub, 2020. https://github.com/eirikbloodaxe/franz/releases/tag/v1.0.
[9] L. Breiman, *Probability*, Addison-Wesley, Reading, MA, 1973.
[10] G.W. Brier, *Verification of forecasts expressed in terms of probabilities*, Monthly Weather Rev. 78 (1950), pp. 1–3.
[11] A.P. Dawid, *Statistical theory: the prequential approach*, J. R. Statist. Soc. Ser. A 147 (1984), pp. 278–292. https://doi.org/10.2307/2981683. MR 763811.
[12] A.P. Dawid and V.G. Vovk, *Prequential probability: principles and properties*, Bernoulli 5 (1999), pp. 125–162. http://dx.doi.org/10.2307/3318616. MR 1673572.
[13] F.X. Diebold and R.S. Mariano, *Comparing predictive accuracy*, J. Bus. Econom. Statist. 20 (2002), pp. 134–144. https://doi.org/10.1198/073500102753410444, Twentieth anniversary commemorative issue. MR 1940633.
[14] M. Ehrendorfer and A.H. Murphy, *Comparative evaluation of weather forecasting systems: sufficiency, quality, and accuracy*, Monthly Weather Rev. 116 (1988), pp. 1757–1770.

[15] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2, John WIley & Sons Inc., New York, 1970.
[16] C.A. Ferro and D.B. Stephenson, *Deterministic forecasts of extreme events and warnings*, in Jolliffe and Stephenson [24], chap. 10, pp. 185–201.
[17] R. Giacomini and H. White, *Tests of conditional predictive ability*, Econometrica 74 (2006), pp. 1545–1578.
[18] T. Gneiting, *Making and evaluating point forecasts*, J. Am. Stat. Assoc. 106 (2011), pp. 746–762.
[19] T. Gneiting, F. Balabdaoui, and A.E. Raftery, *Probabilistic forecasts, calibration and sharpness*, J. R. Statist. Soc.: Ser. B (Stat. Methodol.) 69 (2007), pp. 243–268. http://dx.doi.org/10.1111/j.1467-9868.2007.00587.x.
[20] T. Gneiting and A. Raftery, *Strictly proper scoring rules, prediction, and estimation*, J. Am. Stat. Assoc. 102 (2007), pp. 359–378.
[21] I.J. Good, *Rational decisions*, J. R. Statist. Soc. 14 (1952), pp. 107–114.
[22] T.M. Hamill, *Interpretation of rank histograms for verifying ensemble forecasts*, Monthly Weather Rev. 129 (2001), pp. 550–560.
[23] R. Hogan and I. Mason, *Deterministic forecasts of binary events*, in Jolliffe and Stephenson [24], chap. 3, pp. 31–59.
[24] I.T. Jolliffe and D.B. Stephenson, *Forecast Verification; A Practitioner's Guide in Atmospheric Science*, 2nd ed., John Wiley & Sons, Ltd., Chichester, 2012.
[25] R.W. Katz and M. Ehrendorfer, *Bayesian approach to decision making using ensemble weather forecasts*, Weather Forecast. 21 (2006), pp. 220–231.
[26] J.E. Matheson and R.L. Winkler, *Scoring rules for continuous probability distributions*, Manag. Sci. 22 (1976), pp. 1087–1096.
[27] A. Murphy and R.W. Katz, *Economic Value of Weather and Climate Forecasts*, Cambridge University Press, Cambridge, 1997.
[28] A.H. Murphy and R.L. Winkler, *Reliability of subjective probability forecasts of precipitation and temperature*, J R Stat Soc Ser C Appl. 26 (1977), pp. 41–47.
[29] N. Nolde and J.F. Ziegel, *Elicitability and backtesting: perspectives for banking regulation*, Ann. Appl. Stat. 11 (2017), pp. 1833–1874. https://doi.org/10.1214/17-AOAS1041. MR 3743276.
[30] C. Primo, C.A.T. Ferro, I.T. Jolliffe, and D.B. Stephenson, *Calibration of probabilistic forecasts of binary events*, Monthly Weather Rev. 137 (2009), pp. 1142–1149. http://journals.ametsoc.org/doi/abs/10.1175/2008MWR2579.1.
[31] D.S. Richardson, *Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size*, Q. J. R. Meteorol. Soc. 127 (2001), pp. 2473–2489. http://dx.doi.org/10.1002/qj.49712757715.
[32] D.S. Richardson, *Economic value and skill*, in Jolliffe and Stephenson [24], chap. 9, pp. 165–187.
[33] M. Rosenblatt, *Remarks on a multivariate transformation*, Ann. Math. Stat. 23 (1952), pp. 470–472. MR 0049525.
[34] F. Seillier-Moiseiwitsch and A.P. Dawid, *On testing the validity of sequential probability forecasts*, J. Am. Stat. Assoc. 88 (1993), pp. 355–359. MR 1212496.
[35] F. Seillier-Moiseiwitsch, T.J. Sweeting, and A.P. Dawid, *Prequential tests of model fit*, Scand. J. Stat. 19 (1992), pp. 45–60. MR 1172966.
[36] S. Siegert, O. Bellprat, M. Ménégoz, D.B. Stephenson, and F.J. Doblas-Reyes, *Detecting improvements in forecast correlation skill: statistical testing and power analysis*, Monthly Weather Rev. 145 (2017), pp. 437–450.
[37] S. Siegert, J. Bröcker, and H. Kantz, *Rank histograms of stratified monte-carlo ensembles*, Q. J. R. Meteorol. Soc. 140 (2012), pp. 1558–1571.
[38] I. Steinwart, C. Pasin, R. Williamson, and S. Zhang, *Elicitation and Identification of Properties*, Proceedings of The 27th Conference on Learning Theory, M. Balcan, V. Feldman, and C. Szepesvári, eds., Proceedings of Machine Learning Research Vol. 35, 13–15 Jun, Barcelona, Spain. PMLR, 2014, pp. 482–526. http://proceedings.mlr.press/v35/steinwart14.html.
[39] D.B. Stephenson, B. Casati, C.A.T. Ferro, and C.A. Wilson, *The extreme dependency score: a non-vanishing measure for forecasts of rare events*, Meteorol. Appl. 15 (2008), pp. 41–50. http://dx.doi.org/10.1002/met.53.

[40] O. Talagrand, R. Vautard, and B. Strauss, *Evaluation of probabilistic prediction systems*, in *Workshop on Predictability*. European Centre for Medium Range Weather Forecasts (ECMWF), Reading, United Kingdom, 1997, pp. 1–25.
[41] A.W. van der Vaart, *Time series*, 2010. Lecture Notes.
[42] A.P. Weigel, *Verification of Ensemble Forecasts*, in Jolliffe and Stephenson [24], Chap. 9, pp. 141–166.
[43] D.S. Wilks, *Sampling distributions of the Brier score and Brier skill score under serial dependence*, Q. J. R. Meteorol. Soc. 136 (2010), pp. 2109–2118. http://dx.doi.org/10.1002/qj.709.
[44] D.S. Wilks, *Statistical methods in the atmospheric sciences*, 2nd ed., International Geophysics Series Vol. 59, Academic Press, Oxford, 2006.