Testing the Tests: What Are the Impacts of Incorrect Assumptions When Applying Confidence Intervals or Hypothesis Tests to Compare Competing Forecasts?

ERIC GILLELAND, AMANDA S. HERING, TRESSA L. FOWLER, AND BARBARA G. BROWN

Research Applications Laboratory, National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 16 October 2017, in final form 28 March 2018)

ABSTRACT

Which of two competing continuous forecasts is better? This question is often asked in forecast verification, as well as climate model evaluation. Traditional statistical tests seem to be well suited to the task of providing an answer. However, most such tests do not account for some of the special underlying circumstances that are prevalent in this domain. For example, model output is seldom independent in time, and the models being compared are geared to predicting the same state of the atmosphere, and thus they could be contemporaneously correlated with each other. These types of violations of the assumptions of independence required for most statistical tests can greatly impact the accuracy and power of these tests. Here, this effect is examined on simulated series for many common testing procedures, including two-sample and paired t and normal approximation z tests, the z test with a first-order variance inflation factor applied, and the newer Hering–Genton (HG) test, as well as several bootstrap methods. While it is known how most of these tests will behave in the face of temporal dependence, it is less clear how contemporaneous correlation will affect them. Moreover, it is worthwhile knowing just how badly the tests can fail so that if they are applied, reasonable conclusions can be drawn. It is found that the HG test is the most robust to both temporal dependence and contemporaneous correlation, as well as the specific type and strength of temporal dependence. Bootstrap procedures that account for temporal dependence stand up well to contemporaneous correlation and temporal dependence, but require large sample sizes to be accurate.

1. Introduction

Weather, climate, and hydrologic forecasters and other users need to be knowledgeable about the capabilities and limitations of different numerical weather prediction (NWP) models. Model developers must decide whether recent updates to a model have resulted in improved performance. Administrators often wish to know which forecast system is best. For all these reasons, common statistical measures have been used to summarize forecast performance for over a century. The focus of the current paper is on continuous forecast variables and continuous verification measures, rather than on categorical verification, as well as the specific type and strength of temporal dependence. Bootstrap procedures that account for temporal dependence stand up well to contemporaneous correlation and temporal dependence, but require large sample sizes to be accurate.

Often, competing NWP models are compared by examining individual error statistics for each and declaring the one with the best value of the statistic as the best model (e.g., Liu et al. 2014). Statistical uncertainty is not incorporated into this approach, but a point estimate is merely formed based on a specific sample, when in reality, a distribution of values is possible over all possible samples. To more accurately make inferences about the true performance of an NWP model, the uncertainty in the estimate should be included. More recently, some model comparisons have included estimates of sampling uncertainty through confidence intervals on verification statistics (e.g., Im et al. 2006; Wolff et al. 2014). Thus, users can more accurately assess the statistical significance of any differences they see in standard verification measures. The intervals also explicitly communicate the uncertainty in verification statistics and enforce that each is an estimate, rather than a measurement.

Publisher’s Note: This article was updated on 16 August 2018 to correct the first footnote in the appendix, which contained an error when originally published.

DOI: 10.1175/MWR-D-17-0295.1

© 2018 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).
Though many verification statistics can be compared via parametric confidence intervals, other statistics benefit from a nonparametric interval estimate. In particular, bootstrap resampling methods have been investigated and recommended for many different verification statistics (Wilks 1997). The bootstrap is often seen as a panacea for hypothesis testing that is not burdened with assumptions like those necessary for the parametric tests. However, bootstrap resampling carries its own set of assumptions that need to be checked. For example, the usual procedure assumes that the data are independent and identically distributed (IID), an assumption that is usually violated for most forecast verification analyses. When the sample has temporal dependence, a circular block bootstrap is recommended (Paparoditis and Politis 2001; see section 2b for more details). Once a specific resampling procedure is executed, a sample of the statistic of interest is obtained, and it is from this sample that confidence intervals are derived. Various methods have been proposed for calculating these intervals, and each has its own set of properties.

Model comparisons are best completed using an identical set of cases. This matched comparative verification reduces the dimensionality of the problem (Murphy 1991) and increases the statistical power of the test when the matches are positively correlated. In particular, it allows for paired tests of the significance of the difference between two statistics while eliminating intracase variability. Compared to using confidence intervals on separate statistics for two models, the single interval around the difference provides more power to detect differences and will often give a different result than the use of two intervals (Jolliffe 2007).

Samples of forecasts are rarely independent of one another, particularly when nearby in space or time. In this case, standard confidence intervals are too narrow, as the observations are effectively partial repeats of other observations, resulting in a smaller effective sample size. To account for some of the dependence, the variance inflation factor (VIF; Kutner et al. 2004) has been applied to the variance estimate that is incorporated into the confidence interval calculation. The result is an increase in the interval width to reflect a more appropriate level of uncertainty.

Diebold and Mariano (1995) first introduced a test (denoted DM) to compare the accuracy of two competing sets of forecasts, whereby a weighted average of the empirical autocovariance function is used to obtain an estimate of the variance that accounts for the temporal dependence. This test was not intended to be the only method used to compare statistical models because there are optimal finite-sample tests that can be used to compare model fits, such as $F$ tests for comparing nested Gaussian linear models or model selection criteria such as the Bayesian information criteria (BIC; Schwarz 1978). Indeed, Hansen et al. (2011) introduce a set of tests wherein many pairwise comparisons among the BICs of models fit to the full sample can be made. However, the DM test is still appropriate in many situations. When forecasts are not produced from a statistical model such that the parameters controlling the model fit cannot be identified, then tests that do not involve a likelihood, such as the DM test, are the only options available to compare forecasts. Examples of such forecasts include NWP models, expert forecasts, or cases when the model is proprietary and unknown. In addition, most full-sample, model-comparison procedures rely heavily on one-step-ahead quadratic loss; however, other asymmetric loss functions are oftentimes of primary interest (Hering and Genton 2010, 2011; Gilleland 2013). Diebold (2015) gives a nice review of instances in which the DM test is appropriate and when it is not.

The alternative to full-sample model comparisons is to do out-of-sample forecasting, wherein the data are split into a training and testing set: a statistical model is fit on the training set, and forecasts are produced for the testing set. While full-sample model comparisons are more powerful than out-of-sample testing procedures (Hansen and Timmermann 2015), for a fixed number of parameters, model selection criteria will always select the most heavily mined model. On the other hand, by performing out-of-sample forecasting, it becomes more difficult to overfit the model on the training data, as long as “strategic data mining” in the choice of the split between training and testing sets can be avoided. Thus, even if the models are known a priori and if quadratic loss is used to summarize forecasts, DM-type forecast comparison tests can still be appropriate. In this vein, West (2006) and Clark and McCracken (2013) make assumptions about the models themselves and propose more accurate tests for finite samples, but ultimately, the asymptotic normality assumption of the DM test is still reliable even in this setting (Diebold 2015).

Given the suite of testing procedures, it is of interest to know which tests are the easiest to implement and which are the most accurate. While substantial work has already been carried out in various fields to determine these characteristics, there are special circumstances in weather forecast verification and climate model evaluation that have received less attention. Of particular interest in the present work is the testing of competing forecast models against the same observation series. In this setting, contemporaneous correlation (denoted by $\rho$ herein), where the two forecast models are correlated with each other, is an important attribute to consider (cf. Hering and Genton 2011; DelSole and Tippett 2014).
To see why contemporaneous correlation is an issue in the comparative forecast verification setting, see the appendix, where it is shown that this correlation affects the variance of the loss differential series. Additionally, the strength and type of dependence may affect results for many of the approaches, even those that account for dependence. For example, the VIF assumes a simple autoregressive structure, and the estimate is made based on the order of the autoregressive process. In many cases, an order-1 process is assumed for simplicity, so it is of interest to see how such a test may be impacted if the order is higher.

Hering and Genton (2011) propose a modification to the test introduced by Diebold and Mariano (1995), and they test their approach against the latter under varying sample sizes, strengths of dependence, and strengths of contemporaneous correlation. This work follows the same path using a more comprehensive set of simulation scenarios and also applies several more commonly used tests in weather forecast verification studies. Each test requires a different set of assumptions that may not be met for a particular scenario, and it is important to understand how the tests may be affected in these cases. For example, one test accounts for temporal dependence but assumes a particular model for that dependence. Furthermore, in testing for power, Hering and Genton (2011) consider the case of different means of two error series. Often, weather forecasts are calibrated to have the same means in weather forecast verification. Therefore, we test for power when the means are both zero, but the variability of one error series is higher than the other. It is hoped that the results of this work will help advise users about which methods are most robust to the types of series that are often tested in the forecast verification setting.

2. Review of hypothesis testing and confidence intervals

Wilks (1997, 2011), Hamill (1999), Jolliffe and Stephenson (2011), Jolliffe (2007), and Gilleland (2010) all give considerable background on hypothesis testing and confidence intervals. A brief review for completeness is given here, as well as to establish notation and terminology that will be used.

Suppose interest is in whether or not the parameter $\xi = 0$, where, for example, $\xi$ may relate to the differences in mean error (ME) between two forecasts (simple loss). One might obtain an estimate of $\xi$ that is close to zero, say, $\hat{\xi} = -0.25$. But this value is merely one realization that happened to occur. If the test were rerun on a new set of forecasts, perhaps a positive value would appear. In general, only one realization is obtained, so the test must rely on an assumption about the likelihood of observing the value of the statistic that is realized. To run a hypothesis test, a test statistic $S$ must be calculated that incorporates the uncertainty information along with a direct estimate of the parameter. For example, a typical test statistic is

$$S = \frac{\hat{\xi} - \xi_0}{\text{se}(\hat{\xi})}, \quad (1)$$

where $\hat{\xi}$ is the estimated value of $\xi$, and the standard error $\text{se}(\hat{\xi})$ represents the uncertainty of the parameter estimate, which itself needs to be estimated. For the null hypothesis ($H_0$) of interest, here, $\xi_0$ is set to be zero. If interest were in testing whether or not the true value for $\xi_0$ is 1, then $\xi_0 = 1$ would be used in Eq. (1). Importantly, the distribution of a test statistic must be free of any unknown parameters. For example, the normal approximation ($z$ test) assumes $S \sim N(0, 1)$, which is completely specified without any additional unknown parameters.

Note that a statistical hypothesis test does not answer the question of whether or not the value $\hat{\xi} = -0.25$ is substantial enough as to be important or practically significant, but rather, it points to how likely it is to obtain the specific realization, $\xi$, or any other outcome even less favorable to $H_0$, under the scenario that $H_0$ (e.g., $\xi_0 = 0$) is true. If $\hat{\xi}$ is unlikely (likely) to be realized under $H_0$, then the test should reject (fail to reject) $H_0$.

There are two ways in which the decision based on the test result can be wrong. The decision could be to reject $H_0$ when it is actually true or, conversely, could fail to reject $H_0$ when it is false. The former case is sometimes called a type I error, and the probability of such an error is called the size of the test. Typically, the notation $\Pr(\text{type I error}) = \Pr(\text{reject } H_0|H_0 \text{ is true}) = \alpha$ is used. Hypothesis tests are controlled to keep $\alpha$ low, where $\alpha$ is referred to as the significance level for the test. The second type of error is typically handled by trying to maximize $1 - \Pr(\text{type II error}) = 1 - \beta$, which is referred to as the power of the test and is the probability of detecting a difference that truly exists.

Considering the test statistic $S$ in Eq. (1) again, the probability of observing a test statistic at least as large in absolute value as the one observed, assuming that $H_0$ is true, is found by assuming a specific distribution for $S$. This probability is called the $p$ value of the test. If the $p$ value is less than the size or significance level of the test, then $H_0$ is rejected because it is considered extremely unlikely to observe the test statistic if $H_0$ were true. Of course, the test is expected to give the wrong answer $\alpha \times 100\%$ of the time. Generally, if the $p$ value is nearly 0 (1), one can be comfortable in rejecting (failing to
reject) $H_0$. When the $p$ value is modest, say, 0.17, then the situation is less clear, but it allows for a conversation about the likelihood of a difference that can be used as evidence, even if no concrete decision can be made. The choice of significance level is one that must be made by the experimenter a priori, and it should take into consideration not just the size of the test, but also the power. In many situations relevant to testing for a slightly modified new weather forecast model, the size of the test should be increased in order to increase the power; although, if the sample size is large enough, then smaller test sizes may be reasonable. The usual choices, such as 0.01 and 0.05, were originally made before the advent of computers in order to reduce the number of probability distribution tables that needed to be made (Lehmann 1986, p. 69). For a recent, detailed description of the $p$ value, and for the American Statistical Association’s statement on $p$ values, see Wasserstein and Lazar (2016).

A confidence interval (CI) is considerably more useful than a hypothesis test. Hypothesis tests must be made for any test about the true value of a statistic. For example, if after testing whether or not $\zeta = 0$, it is desired to test $\zeta = 1$, and then a new test must be constructed. With a CI, if one wants to conduct a hypothesis test for $\zeta = 0$, then one need only determine whether or not zero lies within the limits of the interval, and a $(1 - \alpha) \times 100\%$ CI is equivalent to a two-sided $\alpha$-level significance test. To test for a different value, the same CI can be used. Moreover, the length of the interval gives a good impression about the amount of uncertainty in the data that a hypothesis test does not as readily convey.

One drawback of the hypothesis testing and CIs described above is that the true parameter of interest is assumed to be fixed but unknown. This paradigm leads to the somewhat awkward interpretation of $S$ and CI. That is, a $p$ value is not a probability that the true value $\zeta$ is equal to the value of interest, but rather a probability statement about the statistic $S$. Similarly, a CI is not an interval that has a $(1 - \alpha) \times 100\%$ probability that $\zeta$ lies in the interval. Instead, the interpretation of a CI is that if the experiment were repeated 100 times, then one can expect that the true parameter lies within $(1 - \alpha) \times 100\%$ of the intervals.

The approach described here is known as the frequentist approach, and it is possible to obtain similar tests that have a more satisfying interpretation at the expense of further assumptions. In particular, the Bayesian approach assumes that the parameter of interest itself is a random variable whose distribution must be assumed a priori. Information from observed data is used to update knowledge about the parameter’s distribution using Bayes’s formula to arrive at what is known as the posterior distribution. From this posterior distribution, a wealth of information concerning the unknown parameter can be gleaned. The approach can be very useful, provided the prior distribution is reasonable or that the posterior distribution is insensitive to the prior choice. Fiducial inference, which was introduced by Fisher (1934), is a frequentist approach that uses an invariance argument to establish a result that is operationally equivalent to Bayesian inference, but justified by what we would now call a conditional inference argument. However, that argument applied only to location-scale families, and Fisher spent much of the rest of his career trying to show that the same reasoning would apply in more general parametric families (R. L. Smith 2008, personal communication). More recently, interest in the fiducial framework has been rekindled, and rigorous frequentist properties have been established for wide classes of fiducial procedures that put them on a sound mathematical basis from a frequentist perspective (e.g., Hannig 2009; Lidong et al. 2008; Hannig et al. 2006, 2007; Wandler and Hannig 2012). The focus of this paper, however, is on the more commonly used frequentist tests only.

a. Loss functions

Here, three loss functions are considered for summarizing forecast accuracy: simple-, absolute- (AE), and square-error (SE) loss; the result amounts to testing the differences in mean error, mean absolute error (MAE), and mean square error (MSE) between the forecast and observed series. The series of simple, AE, or SE differences are called the loss differential series.

With simple loss, it is possible for a series with, for example, larger errors to test out as being better than a series with smaller errors because of the potential for positive errors to cancel out with negative ones. In this regard, AE and SE loss generally provides more useful tests, but ultimately the summary of interest will depend on the users’ needs.

b. Review of some specific testing procedures

Considering Eq. (1) again, recall that a distribution for $S$ must be assumed in applying standard parametric tests. The different tests considered are tantamount to making different assumptions about the distribution of $S$, and some differ in how the standard error of $\zeta$ is estimated. An exception to this rule is that two types of test statistics might be considered in the context of comparing two different forecast models to the same observed series. One is known as a two-sample test, and the other is a paired test. The two-sample test is a test on the difference in the sample means of the two error series, whereas a paired test is a test on whether or not the
point-by-point sample differences, on average, are different from zero or not.

The test procedures considered here include (i) paired and two-sample \( t \) tests; (ii) paired and two-sample \( z \) tests; (iii) paired and two-sample \( z \) tests with AR(1) VIF applied; (iv) paired bootstrap procedures under varying assumptions; and (v) the Hering–Genton (HG) test (Hering and Genton 2011), which are all paired tests.

1) NORMAL AND STUDENT’S \( T \) TESTS

Rationale for the normal distribution comes from the central limit theorem, which provides justification for assuming a normal distribution for means of random variables. For smaller sample sizes, scaling issues with the variance often lead to violations in the normal assumption, and the Student’s \( t \) test has been shown to be a better approximation. When the sample size reaches about 30, the quantiles of the Student’s \( t \) distribution are approximately equal to those of the normal distribution so that the \( z \) test can generally be applied for sample sizes of 30 or larger. Without VIF, there is an assumption that the sample is realized from an IID population. A violation of this assumption will generally lead to tests with standard error terms that are too small when observations are positively correlated, subsequently leading to larger values of \( \alpha \), which are in turn found to be less likely under \( H_0 \) than they should be. The result is a test that is more likely to incur a type I error even when the sample size is large. With VIF, the variance is inflated through multiplication by a factor based on the strength of dependence.

2) BOOTSTRAPPING

Bootstrap tests do not assume a specific distribution for \( S \) and instead let the data speak for themselves. The primary assumption is that the relationship between the sample distribution and that of the population is consistent with the relationship between the bootstrap resampled distribution and the sample distribution. For example, if \( \hat{\xi} - \xi \sim F \), then \( \hat{\xi}^{*} - \xi \sim F \) also, where \( \hat{\xi}^{*} \) is the bootstrap-estimated parameter value. Nevertheless, assumptions about the distribution of \( S \) are implicit in the bootstrap procedures, and most of the different techniques are aimed at relaxing those assumptions.

Many good reviews of the IID bootstrapping procedures are available (e.g., Davison and Hinkley 1997; Efron and Tibshirani 1998; DiCiccio and Efron 1996). Lahiri (2003) gives an excellent overview of the bootstrap procedure for dependent data, along with a concise, but thorough, review of the IID case. Gilleland (2010) provides an accessible review in the context of forecast verification. While one can calculate a \( p \) value using this paradigm, more often, confidence intervals are desired, and the following discussion uses that terminology.

In perhaps its simplest form, the IID bootstrap procedure is carried out as follows. Given a series of observations \( x = (x_1, \ldots, x_n) \), assume that the realized series is representative of the population series \( X = (X_1, \ldots, X_N) \), and perform the following steps.

1) Draw a sample of size \( n \) with replacement from \( x \), say, \( x^*_b = (x^*_{1(b)}, \ldots, x^*_{n(b)}) \).
2) Estimate the parameter(s) of interest from the resampled data, denoted \( \hat{\xi}^*_b \), using the sample \( x^*_b \).
3) Repeat steps 1 and 2 \( B \) times to obtain a sample of the estimate(s) \( \hat{\xi}^*_B = (\hat{\xi}^*_{1}, \ldots, \hat{\xi}^*_B) \).
4) Calculate confidence limits from the sample obtained in step 3.

The simplest, and perhaps most common, method for obtaining the confidence interval in step 4 is to use the percentile method, whereby the \( \alpha/2 \) and \( 1 - \alpha/2 \) quantiles from \( \hat{\xi}^*_{1}, \ldots, \hat{\xi}^*_B \) are chosen so that the interval is simply \( (\hat{\xi}^*_{B[\alpha/2]}, \hat{\xi}^*_{B[1-\alpha/2]}) \) with the \( [q] \) subscript to denote the \( q \)th quantile. This method is straightforward and works well if \( X \) are IID and the estimator \( \hat{\xi} \) of \( \xi \) is unbiased. In fact, it has been shown that for estimating the mean, the bootstrap interval is more accurate than the normal approximation method (Singh 1981). If the assumptions of unbiasedness and constant variance are not met, then the bias-corrected and adjusted (BCa) method will give highly accurate intervals. Because the percentile method intervals result as a special case of the BCa procedure when the assumptions are met, the BCa method is generally a better choice. However, because the BCa requires an additional set of resampling, it can be computationally expensive to calculate. Therefore, the percentile method is sometimes still preferred. Either interval is range preserving (i.e., it is impossible to obtain limits for \( \xi \) that are not in the natural range of the parameter) and transformation respecting [i.e., if a monotonically increasing function \( f \) is applied to the limits, the resulting interval is appropriate for \( f(\xi) \)].

A third option is the original method, also known as the “basic bootstrap” interval, which is given by \( (\hat{\xi} - \hat{\xi}_{B[1-\alpha/2]}, 2\hat{\xi} - \hat{\xi}_{B[\alpha/2]}) \), where \( \hat{\xi} \) is the estimated statistic obtained using the original dataset.

While the bootstrapping procedure results in useful intervals that can be more accurate than parametric intervals, such as those based on the normal approximation, many of the methods for obtaining the final interval from the sample of test statistics, such as the percentile and BCa, are intuitively unappealing, at least from a mathematical statistics point of view. That is, the estimate \( \hat{\xi} \) is being considered as the true value so that the comparison is one of \( \hat{\xi} - \xi \) instead of \( \hat{\xi} - \xi \). The basic
bootstrap method is a more intuitive method for obtaining the final interval, which can be used for both the IID and the block bootstrap approaches. However, the basic bootstrap interval is neither range preserving nor transformation respecting and, in some instances, is less accurate than the percentile and BCA methods. A modification of the basic bootstrap that is also not range preserving or transformation respecting, but is as accurate as the percentile methods, is the bootstrap serving or transformation respecting, but is as accurate as transformation respecting and, in some instances, is less basic bootstrap interval is neither range preserving nor IID and the block bootstrap approaches. However, the

...the percentile methods, is the bootstrap serving or transformation respecting, but is as accurate as transformation respecting and, in some instances, is less basic bootstrap interval is neither range preserving nor IID and the block bootstrap approaches. However, the

...the percentile methods, is the bootstrap serving or transformation respecting, but is as accurate as transformation respecting and, in some instances, is less basic bootstrap interval is neither range preserving nor IID and the block bootstrap approaches. However, the

The choice for the number of resamples $B$ is made by attempting to find the smallest $B$ for computational efficiency, where a higher value will not change the results substantially. To obtain all possible combinations of a data series of size $n$ would require $2n - 1$ choose $n$ resamples [i.e., $(2n - 1)!/[n!(n - 1)!]]$, which is a very large number even for small $n$ (e.g., if $n = 10$, there are 92378 total combinations). Here, $B = 500$ is used; $B = 1000$ was also tested, but yielded similar results.

3) DM AND HG TESTS

Both the DM and HG tests rely on an assumption that the loss differential is covariance stationary, meaning that the series of random variables $X = (X_1, \ldots, X_N)$ has a constant mean with finite variance, and the covariance between any two random variables is a function of only the distance in time between them. Given this assumption, the test statistic in Eq. (1) is asymptotically standard normal. The difference between the DM and HG tests is in how the standard error of $\hat{\zeta}$, the denominator in Eq. (1), is calculated. For the DM test, a consistent estimator of $\sigma^2(\zeta)$ is obtained by taking a weighted sum of the sample autocovariance function (ACF) for a $k$-step-ahead forecast. These sample autocovariances are computed for each lag $\tau$ as

$$\hat{\gamma}_3(\tau) = \frac{1}{N} \sum_{t=|\tau|+1}^N (x_t - \bar{x})(x_{t-|\tau|} - \bar{x}),$$

but in practice, their estimator sometimes results in a negative value, and the sum over all lags of the empirical ACF is identically zero. The HG estimator avoids these problems by fitting a parametric covariance model to the empirical ACF that is guaranteed to be positive definite. In practice, an exponential model fits many empirical ACFs very well, but other choices can be made depending upon the structure of the ACF. To obtain the most appropriate parametric model, the model is fit only to lag correlations that are shorter than about half the maximum lag for the observed series, while the final estimate of the standard error is summed over all lags from the modeled ACF.

3. Simulations

To gauge the effect of contemporaneous correlation on the hypothesis testing procedures, as well as differing serial dependence structures, two types of simulations are considered. The first, described in section 3a, induces serial dependence by way of a simple moving average model that makes it easy to vary the strength of the dependence while also varying the contemporaneous
correlation between two series. Beyond this exercise, differing dependence structures are examined between uncorrelated simulated error series in order to evaluate whether, and if so, to what extent, certain dependence structures affect each testing procedure. Section 3b describes simulations created for this purpose.

a. MA(1) simulations

To compare results here with those from Hering and Genton (2011), the same first-order moving average [MA(1)] simulation study is conducted, which varies the sample size, range of temporal dependence, and contemporaneous correlation of two sets of time series. Hering and Genton (2011) found that the HG test was superior to the DM test in size and that the HG test is a powerful test. However, in their testing procedure for power, they varied the mean and not the variance of the second error series. Because weather forecasts are often calibrated prior to verification studies, it makes sense to study power in this setting by varying the variance of the second series. Therefore, power will be tested again, under this paradigm, for the HG test. Figure 1 (top row) shows an example of a simulated MA(1) series along with its ACF and PACF plots.

Fig. 1. (left) Example time series following (top) an MA(1), (middle) AR(2), and (bottom) ARMA(1, 1), along with the (center) associated ACF and (right) PACF plots.
b. Other dependence structures

In addition to the above simulations, a few different temporal dependence structures are considered: an AR(2) and two autoregressive moving average (ARMA) processes. An ARMA(p, q) process is defined by

\[ X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}, \]

where \( \epsilon_t \sim N(0, \sigma^2) \). For studying empirical size, \( \sigma = 1 \) for both simulated error processes. When studying power, the standard deviation for the first series will remain as \( \sigma_1 = 1 \), and the standard deviation for the second series \( \sigma_2 \) will vary to be 1.5, 2, 2.5, and 3. If a test is powerful for \( \sigma_2 = 1.5 \), then there is no need to test it for higher values. An AR(1) [AR(2)] series is the same as an ARMA(1, 0) [ARMA(2, 0)] series.

As an example, the AR(1) series, which is the assumed structure for the \( z \) test with VIF used herein, is given by

\[ X_t = \phi X_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2). \quad (3) \]

Here, the three time series models considered are

(i) AR(2): \( X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t, \)
(ii) ARMA(1, 1): \( X_t = \phi X_{t-1} + \epsilon_t + \theta \epsilon_{t-1}, \)
(iii) ARMA(2, 2): \( X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}, \)

where, again, the residual errors, \( \epsilon_t \sim N(0, \sigma^2) \). Examples of (i) and (ii) are shown in Fig. 1 (middle and bottom rows, respectively).

Simulations from the AR(2) process have variance (Hamilton 1994, p. 58):

\[ \text{Var}(X_{it}) = \frac{(1 - \phi_1 \phi_2) \sigma^2}{(1 + \phi_2)[(1 - \phi_2)^2 - \phi_1^2]} \quad i = 1, 2. \]

Therefore, each simulated series is multiplied by \( \{(1 - \phi_2)/(1 + \phi_2)[(1 - \phi_2)^2 - \phi_1^2]\}^{1/2} \) in order to ensure that the variance is simply \( \sigma^2 \), the innovation variance for the given series. All series are simulated to have unit variance, except when checking for power, where the second series will have variances \( \sigma_2^2 = 1.5^2, 2^2, 2.5^2, \) or \( 3^2 \).

The ARMA(1, 1) processes have variance (cf. Brockwell and Davis 2010, chapter 3):

\[ \text{Var}(X_{it}) = \sigma_i^2 \left[ 1 + \frac{(\theta + \phi)^2}{1 - \phi^2} \right], \quad i = 1, 2, \]

so that simulations are multiplied by \( [1 + (\theta + \phi)^2]/(1 - \phi^2)]^{1/2} \) in order to ensure they have variance \( \sigma_i^2 \) for each of the 1000 simulated series.

Obtaining the necessary multiplication factor for the ARMA(2, 2) process is considerably more complicated to spell out, but it is easily found using, for example, one of the methods described in Brockwell and Davis (2010, chapter 3), and for the specific values used in this study, it is about one-third.

The AR(2) and ARMA simulations are carried out using the arima.sim function from the R (R Core Team 2014) package stats. The parameters used for the simulated series are the following:

- AR(2): \( \phi_1 = 0.8, \quad \phi_2 = -0.2 \)
- ARMA(1,1): \( \phi = 0.8, \quad \theta = -0.3 \)
- ARMA(2,2): \( \phi_1 = 0.8, \quad \phi_2 = -0.4, \quad \theta_1 = -0.2, \quad \theta_2 = 0.25 \)

All simulations are performed using 1000 replications. For example, if the sample size under study is 100, then 1000 samples of size 100 are simulated. This number is chosen arbitrarily but should be sufficiently large as to yield robust results.

Because the paired tests are conducted on the loss differentials, one might think that perhaps the temporal dependence is removed through the differencing. However, such dependence is not removed and generally takes on a form that is very similar to the two series being differenced; when the two series have different dependence structures, the result is generally a series with a combination of the structures. For a discussion on properties of the loss differential series, including the resulting dependence structures, see the appendix. The \( z \) test with VIF applied assumes a particular \( AR(p) \) dependence model, and most often the simplest choice with \( p = 1 \) is employed, as is carried out here. These models with these parameter choices have dependence structures that the VIF based on the AR(1) model does not take into account.

4. Results

In this section, the MA(1) model described in section 3a is simulated to have contemporaneous correlation of \( \rho = 0, 0.5, \) and 0.9, as well as varying the temporal dependence range parameter \( \theta = 0, 0.5, \) and 0.9 for sample sizes \( n = 8, 16, 32, 64, 128, 256, \) and 512, which results in \( 3 \times 3 \times 7 = 63 \) combinations of parameters and sample size. The AR(2), ARMA(1,1) and ARMA(2,2) models are also simulated, as described in section 3b.

Each simulation is performed 1000 times for each combination, and each of the seven tests (two-sample and paired \( t \) tests, normal approximation \( z \) test with and without VIF applied, the IID and CB bootstrap methods, and the HG test) with significance level \( \alpha \) are applied, yielding a binary sample of 1000 giving 0 if \( H_0 \) is not rejected and 1 if it is. For testing size,
\(H_0\) is simulated to be true so that the correct answer is 0, and it is expected that \(\alpha\%\) of the 1000 tests will be 1. The actual percentage of times a test results in 1 gives the empirical size for the testing procedure, which should be close to \(\alpha\) if the test is accurate. Empirical size results are given as percentages, and the closer the value is to \(\alpha\), the better the test. For testing empirical power, \(H_0\) is not the true distribution, so the higher the percentage of 1-valued tests, the better.

In terms of size, the level of the test is generally found to not be an important factor, so only the 10\% level results are displayed for brevity, except when a marked difference does occur. For tests that have appropriate size and are competitive with the HG test, power is also tested empirically. Otherwise, if a test does not have appropriate size, it should not be considered further. Furthermore, results are only shown for AE loss, as they are similar to those for SE loss. Comments are made when differences do occur.

a. Size

1) MA(1) Simulations

To test for the effects of sample size, range of temporal dependence, and contemporaneous correlation, the MA(1) simulations conducted in Hering and Genton (2011) for the DM and HG tests are employed. Hering and Genton (2011) already determined that the HG test has good empirical size using these simulations, better than the DM test. Therefore, it remains to compare the other, more commonly used tests that were not compared by Hering and Genton (2011).

Figures 2–4 show the empirical size results for the MA(1) simulations, with the HG test results included for reference. The plots give the results using AE loss. Those for SE loss are found to be analogous, and the simple loss results mostly show a distinct lack of accuracy for tests, except for HG, where the results are analogous to those for AE and SE. Results for both IID and CB bootstrap methods are shown for the basic CI method. Results for the percentile method are analogous, while those for BCa are similar, but the accuracy is more variable. The closer a test’s empirical size is to the horizontal dotted line (through \(\alpha = 10\%\)), the more accurate the specific testing procedure is. Values below this line mean that the test is undersized (rejects \(H_0\) less often than it should), and values above mean that the test is oversized (rejects \(H_0\) more often than it should).

When no temporal dependence and no contemporaneous correlation are present, all of the tests fare similarly in terms of accuracy, although the CB bootstrap is oversized by a larger amount than all other tests, becoming less oversized to about 2\% for larger sample sizes, which suggests that its accuracy is diminished when temporal dependence is not present (cf. Fig. 2, top-left panel). With the introduction of moderate contemporaneous correlation (\(\rho = 0.5\); Fig. 2, top-right panel), the two-sample \(t\) and both \(z\) tests become severely undersized, with an empirical size estimated at about half the value of \(\alpha\). With strong contemporaneous correlation (\(\rho = 0.9\); Fig. 2, bottom-left panel), the empirical size for these tests is zero. The IID and CB bootstrap, paired \(t\), and HG test do not appear to be greatly affected by the addition of contemporaneous correlation without temporal dependence, even stronger such correlation. Indeed, both the \(t\) test and IID bootstrap are competitive with the HG test under contemporaneous correlation, provided no temporal dependence is present.

Although the paired \(t\) test fared well for accuracy in the face of contemporary correlation with no temporal dependence, the addition of moderate temporal dependence (\(\theta = 0.5\); Fig. 3) causes the test to become oversized by about 50\% of \(\alpha\). The behavior of the \(z\) test without VIF applied is erratic in this case, as it is oversized by about 50\% with no contemporaneous correlation (Fig. 3, top-left panel), has relatively good accuracy when contemporaneous correlation and temporal dependence are moderate (both at 0.5; Fig. 3, top-right panel), and has empirical size that is nearly zero in the case with strong contemporaneous correlation and moderate temporal dependence. Similarly, for simple loss (not shown), the only scenario where the \(z\) test and two-sample \(t\) test are accurate occurs when the contemporaneous correlation is set to \(\rho = 0.5\) and the temporal dependence to \(\theta = 0.9\). The IID bootstrap is oversized, but the CB bootstrap does not appear to be affected any further for these cases, and the HG test is very accurate. The tests worsen similarly in terms of accuracy as the temporal dependence is increased, except for the CB bootstrap, which seems to improve, and the HG test, which is not affected.

The \(z\) test with VIF applied is grossly inadequate for these MA(1) simulations, having nearly zero test size for all but the case of no correlation, where it is reasonable, and under moderate contemporaneous correlation without temporal dependence, where it is still undersized by about half of \(\alpha\).

Regardless of the scenario, none of the tests are very accurate for small sample sizes, except in the case where no type of dependence is introduced, where both \(t\) tests are the most accurate for the two smallest sample sizes (\(n = 8, 16\)). Under SE loss, the \(t\) test still performs well for the smallest sample size tested of \(n = 8\), but it is found to be undersized by about 2\%. The bootstrap tests require larger sample sizes, as they do not achieve reasonable accuracy until the sample sizes reach about \(n = 250\).
The poor results for the z test + VIF method can be explained by the fact that the simulations are from an MA(1) rather than an AR(1) process, upon which the VIF applied here is based. While both series have a constant mean of zero and constant variances, they have very different correlation structures, and the VIF utilizes the correlation structure of the AR(1) to inflate the variance. In particular, an AR(1) series has correlation \( \text{Corr}(X_t, X_{t+h}) = \phi^h \), whereas an MA(1) process has correlation \( \text{Corr}(X_t, X_{t+h}) = \theta/(1 + \theta^2) \) for \( |h| \leq 1 \) and 0 otherwise. For 1000 simulations of the MA(1) processes with contemporaneous correlation \( \rho = 0 \) (none) and \( \rho = 0.5 \) (moderate) and \( \theta = 0.5 \), empirically estimated AR(1) parameters from the resulting AE loss series \( \hat{\phi} \) are found to fall between 0.6 and 0.75 for either value of \( \rho \). This result leads to an estimated AR(1) VIF, \( (1 + \hat{\phi})/(1 - \hat{\phi}) \), in the range of 4.5 to 6.5, which, for the case with \( \rho = 0 \), results in an estimated variance that is rather large and likely to be too conservative, as the empirical size tests show.

2) OTHER DEPENDENCE STRUCTURES

(i) Parametric tests

Empirical size results (not shown) for the AR(2) simulations using the parametric testing procedures that assume IID data do not fare very well. They are substantially oversized, especially under simple loss. On
the other hand, tests that apply the VIF perhaps overcompensate because they are very undersized, and it is found that their power is also very low. Results for empirical size using the ARMA(1, 1) and ARMA(2, 2) models are analogous to those for the AR(2).

(ii) Bootstrap tests

Empirical size results (not shown) for the AR(2), ARMA(1, 1) and ARMA(2, 2) simulations show that because of the strong dependence in each of the simulated series, the IID bootstrap tests are oversized by a larger magnitude than the other tests. The CB bootstrap method yields tests that are rather oversized for very small sample sizes, but only oversized by a small amount as the sample size increases beyond about 100, and it approaches the correct size with increasing sample sizes. The type of dependence structure does not appear to affect the CB method, which is both desired and expected. Moreover, it clearly outperforms the parametric test with VIF applied in terms of size. It is reassuring that the CB bootstrap approach performs well under these scenarios.

b. Power

As described in section 3, when checking for power, the assumption is that any bias will have been already removed from the forecast via calibration so that only the error variances are of interest here. In particular, while checking size properties under simple loss made sense, it is not possible to test for power under this paradigm for this loss function because $H_0$ remains true, even if the error variances differ, because $E(u_1) - E(u_2) = 0 - 0 = 0$ even if $\sigma_2$ is much larger than $\sigma_1$. Subsequently, power is only examined for AE and

---

**Fig. 3.** As in Fig. 2, but for moderate temporal dependence range ($\theta = 0.5$).
SE loss where the null hypotheses are $E(|u_1|) = E(|u_2|) = 0$ and $E(u_1^2) - E(u_2^2) = 0$, respectively. Because the means are zero, the difference being tested under SE loss is $\sigma_1^2 - \sigma_2^2$ because for a random variable $X$ with mean $\mu$ and variance $\sigma^2$, $\text{Var}(X) = E(X^2) - E^2X = \sigma^2 - \mu^2$. Again, unless results differ drastically for SE, only AE results are shown.

1) MA(1) SIMULATIONS

Empirical power results are carried out for the case of $\theta = 0$ using $\sigma_2^2 = 1.5^2\sigma_1^2$ for the paired $t$ test (not shown), as it is found to have accurate empirical test size in this setting. A higher percentage of rejected tests is better because under the power simulations, the mean loss differential does not follow $H_0$; generally, at least 80% is sufficiently high. The test is primarily only powerful for larger sample sizes ($n \geq 60$), and stronger contemporaneous correlation increases the power. When $\sigma_2^2 = 2^2\sigma_1^2$, the test is powerful for sample sizes of at least 15, regardless of contemporaneous correlation.

Similar to the paired $t$ test, the IID bootstrap procedure is found to have accurate size, regardless of contemporaneous correlation when no temporal dependence is introduced. The CB bootstrap is found to have reasonable accuracy, regardless of contemporaneous correlation or temporal dependence, but is nevertheless oversized. Therefore, power is tested only for the IID approach. With $\sigma_2^2 = 1.5^2\sigma_1^2$, the IID bootstrap is found to be powerful, with empirical power above 80% when the sample size is at least $n = 60$, regardless of which CI method is used.

The HG test (Table 1) is powerful when $\sigma_2^2 = 1.5\sigma_1^2$ when the sample size is at least 60, regardless of temporal dependence and contemporaneous correlation. However, stronger contemporaneous correlation appears...
to have a positive effect on the power of the HG test, as the empirical power increases consistently with $\rho$ and is above 80% for sample sizes as small as 16 when $\rho = 0.9$. Recall that the empirical size for the HG test is not affected by the contemporaneous correlation. As $\sigma^2_T$ is increased (not shown), the sample size required for powerful tests decreases considerably to 16.

Despite the fact that the CB bootstrap’s empirical size is consistently too high, for larger sample sizes, the amount of its overage is relatively small: on the order of about 2% over $\alpha$. Therefore, its power is also tested here (not shown), and for sample sizes over 30, it is a powerful test.

2) Other Dependence Structures

Bootstrap tests The CB bootstrap method requires larger sample sizes to have any power for the AR(2) simulations. For $\alpha = 0.01$ and $\sigma^2_T = 1.5$, power is generally quite low, reaching only about 65% for sample sizes of 500 under AE loss and slightly higher for SE loss. Smaller sample sizes are possible as $\alpha$ increases, reaching about 65% at $n = 300$. Results are similar for the MA(1) simulations and the AR(2) simulations, although power is a bit lower for $\alpha = 0.01$ and a bit higher for $\alpha = 0.05$. ARMA(2, 2) simulations also yield very similar results.

5. Conclusions

It is often necessary to compare two competing forecasts in terms of the differences in their verification summaries. The main objective of this paper is to test common statistical inference procedures for comparing competing forecast models in the face of both temporal dependence and contemporaneous correlation. Multiple testing procedures for comparing competing forecasts against the same set of observations are examined to determine their accuracy in terms of size and, when appropriate, power. The tests are conducted on simulated series with known types of temporal dependence, as well as with contemporaneous correlation between the competing models. The latter is an often overlooked issue that can have a large impact on results (cf. DelSole and Tippett 2014; Hering and Genton 2011; Li et al. 2016).

The simulated series with contemporaneous correlation are from the same MA(1) simulations as those applied by Hering and Genton (2011), which were used to test their method, the HG test, against the one proposed by Diebold and Mariano (1995). Because Hering and Genton (2011) do not include other more common tests in their work, such testing is conducted here. In the interest of brevity, only the most commonly applied testing procedures are considered along with the HG test. In particular, the usual two-sample and paired $t$ and $z$ tests are carried out both with and without the variance inflation factor (VIF) applied, where the VIF is based on an AR(1) model. Furthermore, the IID and CB bootstrap methods are carried out using various different methods for constructing CIs from the resulting parameter resamples.

Additional simulations from other time series models are also employed in order to ascertain whether or not the specific type of dependence structure affects the outcomes. As shown in the appendix, the loss differential series do not lose any of the dependence from the individual series through the differencing process; rather, they retain a dependence structure that inherits from both series, resulting in a more complex dependence structure. Therefore, the additional simulations shed light on whether the results here are likely to be more general or not.

The appendix illustrates that the series resulting from taking the difference between two AR(1) series with identical autoregressive parameter $\phi$ would result in another AR(1) process with autoregressive parameter $\phi$. However, in practice, it is unlikely that two series will have identical autoregressive parameters, and the result of taking the difference between two such series is no longer an AR(1) process. In fact, a similar exercise to the
MA(1) study described in section 3a was also carried out, but using AR(1) models. The results did not improve. Therefore, one conclusion from this study is that none of the traditional tests, even the $z$ test + VIF, should be used when testing which of two competing forecasts is better.

The HG test is included mainly for power comparisons, where power is determined by varying the variance of the second simulated series instead of the mean because weather forecast verification and climate model evaluation are often calibrated to have the same mean as the observation series. Thus, we have shown that the HG test can also be used to test for differences in variances of two sets of calibrated forecasts. Results for empirical size for the HG test are also shown for comparison and because they were not reported for AE loss in Hering and Genton (2011). Methods that test for equality of variances, lagged autocorrelation, and/or means between two time series do exist, but they often ignore contemporaneous correlation (e.g., Lund et al. 2009; Lund and Li 2009).

It is verified that the paired testing procedures are considerably more accurate than the two-sample tests, which work well only when no contemporaneous correlation or temporal dependence exist and are otherwise fairly undersized. While the paired t test is reasonably accurate for size, regardless of contemporaneous correlation, the test is generally not very powerful, although it still has good power if the sample size is large and no temporal dependence is present. Generally, in the realm of forecast verification, where strong contemporaneous correlation and temporal dependence are likely to occur, the test is not recommended. The accuracy of the paired $z$ test is greatly affected by both contemporaneous correlation and temporal dependence. The $z$ test with VIF applied, where the VIF is calculated assuming an AR(1) model, is not accurate when the temporal dependence structure is not approximately AR(1).

All of the bootstrap procedures are found to be reasonably robust to contemporaneous correlation and structure of temporal dependence, but require fairly large sample sizes before they become accurate. One exception is for the IID bootstrap in the case of no temporal dependence, which is accurate for relatively small sample sizes. For the largest sample size simulated ($n = 512$), the test is a little oversized, even when accounting for temporal dependence. The tests confirm that the IID bootstrap resampling is inappropriate when temporal dependence is present and should generally not be used for forecast verification. The circular block (CB) bootstrap, on the other hand, is fairly accurate, provided the sample size is large. It is less accurate than the IID bootstrap when no temporal dependence is present. The type of CI method used for the bootstrap procedures does not have an impact on the size or power results found in this study.

The HG test is the most accurate, powerful, and robust of all of the procedures compared. Its main drawback is the necessity of fitting a parametric model to the autocovariance function (ACF), which can add computational cost, and automated optimization numerics can sometimes fail. Li et al. (2016) take a functional approach to test for a difference in means and/or variances between two space–time fields wherein sensitivity to model misspecification is reduced, and future work could propose an analogous method for time series. Nevertheless, the HG test is otherwise a straightforward procedure that makes a fairly modest modification to the classical tests, which are the easiest to implement but are also the least appropriate. The CB method requires multiple resampling and, given the need for large samples, can also be computationally expensive, generally much more so than the HG test, which could be implemented on a smaller subset if need be.

It is reasonable to expect that some forecast models might capture the observed series behavior very well, but could have a timing issue so that the direct match might result in a much poorer measure. It is even possible that one forecast might have a small timing error but otherwise be perfect, but the test might prefer a worse model that does not have the timing error. Such issues are not handled by any of the testing procedures studied here, in part because the answer to the question of which model is better becomes more complex, making size and power comparisons difficult to conduct. Further, if timing errors are less important to a user, then the forecast could be calibrated for timing errors before performing the test. Alternatively, such a realignment can be performed within the testing procedure (cf. Gilleland and Roux 2015).

The tests investigated in this study pertain to univariate time series. Hering and Genton (2011) proposed and tested an extension of their testing procedure to the spatial domain and found it to also have appropriate size and good power, provided the spatial and contemporaneous correlations are not too strong. Gilleland (2013) introduced a loss function within their testing procedure that allows for spatial displacement errors to be taken into account. While the HG test in either its univariate or spatial form is relatively new, it is already beginning to be used in practice (e.g., Ekström and Gilleland 2017; Gilleland et al. 2016; Lenzi et al. 2017; Razafindrakoto et al. 2015; Zhang et al. 2015).

Acknowledgments. Support was provided by the National Science Foundation (NSF) through Earth System Modeling (EaSM) Grant AGS-1243030. The National Center for Atmospheric Research is sponsored by the National Science Foundation.
APPENDIX

Properties of the Loss Differential for the Simulations

The loss differential series is a combination of two time series models, which will result in a new series that inherits certain properties from the original series. The specific forms for the loss differential series are not crucial to this study because of the simulation focus. However, it can be instructive to understand how certain test assumptions may be violated in the domain of forecast verification. First, expected values and variances are derived for loss differential series under the much simpler case of IID series before investigating how dependence structures carry over from the original to the loss differential series.

a. IID series

If it is supposed that the observed series, say, $\hat{Y}$, is an estimate of the true state $Y$, then it can be written as

$$E(|\hat{X}_1 - \hat{Y}| - |\hat{X}_2 - \hat{Y}|) = E|\eta_1| - E|\eta_2| = \sqrt{\frac{2}{\pi}} \{\exp[-b_1^2/(2\sigma_1^2)] - \exp[-b_2^2/(2\sigma_2^2)]\}$$

$$+ b_1 \left[1 - 2\Phi\left(-\frac{b_1}{\sigma_1}\right)\right] - b_2 \left[1 - 2\Phi\left(-\frac{b_2}{\sigma_2}\right)\right],$$

which, for unbiased forecasts, reduces to $\sqrt{2/\pi}(\sigma_1 - \sigma_2)$. Thus, even for calibrated forecasts, differences in their variances propagate through to the mean of the loss differential. Similarly, for SE loss, the expected value of the loss differential series under the IID assumption is given by

$$E[(\hat{X}_1 - \hat{Y})^2 - (\hat{X}_2 - \hat{Y})^2] = E\eta_1^2 + E\eta_2^2$$

$$= \sigma_1^2 - \sigma_2^2 + b_1^2 - b_2^2,$$

which is also nonzero for calibrated forecasts with differing variances.

For expected values, contemporaneous correlation $\rho$ is not an issue, but it does appear in the variances of the loss differential. The variance for the IID loss differential series under simple loss is

$$\text{Var}[\hat{X}_1 - \hat{Y} - (\hat{X}_2 - \hat{Y})] = \text{Var}(\hat{X}_1 - \hat{X}_2) = \text{Var}(\eta_1 - \eta_2) = \text{Var}(\eta_1) + \text{Var}(\eta_2) - 2\text{Cov}(\eta_1, \eta_2) = \sigma_1^2 + \sigma_2^2 - 2\sigma_1 \sigma_2 \rho.$$

The variances for the loss differential series with AE and SE losses are considerably more complicated, but both involve the contemporaneous correlation.

b. Dependent series

Under simple loss, it is very straightforward to calculate what sort of time series dependence is present for the loss differential series, but it is also possible to determine the nature of the loss differential under more complicated loss functions. In doing so, it is helpful to use the difference operator $B$, where $BX_t = X_{t-1}$, $B^2X_t = X_{t-2}$, $B^3X_t = X_{t-3}$, and so on. For an AR(1) series, $X_t = \phi X_{t-1} + \epsilon_t$ with $\epsilon_t \sim N(0, \sigma_\epsilon^2)$, can be written $X_t - \phi X_{t-1} = \epsilon_t$ or $(1 - \phi B)X_t = \epsilon_t$. Additionally, the inverse of $(1 - \phi B)$, written $(1 - \phi B)^{-1}$, is equivalent to $(1 + \phi B + \phi B^2 + \phi B^3 + \cdots)$.

The properties below pertain to the simulated series considered in this paper and would generally require complicated loss functions.
certain assumptions that are left out here for brevity. For ease of understanding, it is also shown how the loss differential series would behave as a result of combining two AR(1) processes.

It is easy to verify that under simple error loss, two AR(1) series with identical autoregressive parameters \( \phi \) yield a loss differential series that is also AR(1) with innovation variance \( \sigma^2_1 + \sigma^2_2 + 2\sigma_1\sigma_2\rho \). In general, however, if the two series were to have different autoregressive coefficients, then the loss differential series is ARMA(2, 1). Because it is easy for this case to see how the result is obtained using Eq. (3) directly, this argument is given first, and then again using the difference operator notation. Let \( X_t = \phi X_{t-1} + \epsilon_t \), with \( \epsilon_t \sim N(0, \sigma^2_1) \) and \( W_t = \phi W_{t-1} + \eta_t \), with \( \eta_t \sim N(0, \sigma^2_2) \) represent the two sets of error series. Then, \( X_t - W_t = \phi X_{t-1} - \phi W_{t-1} + \epsilon_t - \eta_t = \phi(X_{t-1} - W_{t-1}) + \lambda_t \), where \( \lambda_t \) is a white-noise process with variance \( \text{Var}(\lambda_t) = \text{Var}(\epsilon_t - \eta_t) = \text{Var}(\epsilon_t) + \text{Var}(\eta_t) + 2\text{Cov}(\epsilon_t, \eta_t) = \sigma^2_1 + \sigma^2_2 + 2\sigma_1\sigma_2\rho \). Using the difference notation, the two series can be written

\[
(1 - \phi B)X_t = \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2_1),
\]
\[
(1 - \phi B)W_t = \eta_t, \quad \eta_t \sim N(0, \sigma^2_2). \tag{A1}
\]

Clearly,

\[
(1 - \phi B)X_t - (1 - \phi B)W_t = (1 - \phi B)(X_t - W_t)
= \epsilon_t - \eta_t. \tag{A2}
\]

For two AR(1) time series, following Hamilton (1994, p. 108), \( X_t \) and \( W_t \), with differing autoregressive coefficients, say, \( \pi \) and \( \kappa \), respectively, can be written in the form

\[
(1 - \pi B)X_t = \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2_1),
\]
\[
(1 - \kappa B)W_t = \eta_t, \quad \eta_t \sim N(0, \sigma^2_2). \tag{A2}
\]

Multiplying the first equation in Eq. (A2) by \( 1 - \kappa B \) and the second by \( 1 - \pi B \) gives

\[
(1 - \pi B)(1 - \kappa B)X_t = (1 - \kappa B)\epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2_1),
\]
\[
(1 - \pi B)(1 - \kappa B)W_t = (1 - \pi B)\eta_t, \quad \eta_t \sim N(0, \sigma^2_2). \tag{A3}
\]

Thus, the difference series \( Y_t = X_t - W_t \) is (1 - \( \pi B \))(1 - \( \kappa B \)(\( X_t - W_t \)) = (1 - \( \kappa B \))\( \epsilon_t \), which is simplified to \( 1 - (\kappa + \pi)B + 2\kappa\pi B^2 \) if \( \pi = (1 - \theta B)\nu_t \) or just \( 1 - (\phi_1 B - \phi_2 B^2)Y_t = (1 - \theta B)\nu_t \), an ARMA(2, 1) process.

Figure A1 shows the ACF\(^1\) [Eq. (2)] and partial autocovariance function (PACF) plots for one set of the AR(1) simulations of two (correlated) series.

\(^1\) Fig. A1 shows the auto-correlation function graph, labeled ACF in the figure, instead of the auto-covariance function (labeled ACF in the text). The auto-correlation function is the auto-covariance function divided by its value at lag zero. The PACF is a conditional correlation of a time series with its own lags controlling for shorter lags (see Brockwell and Davis 2010 for more details). A similar argument can be posed to show that under AE loss, the loss differential for two AR(1) series with the same autocorrelation coefficient is also AR(1). To verify, taking the absolute value of each series in Eq. (A1) gives the loss differential

\[
|X_t| - |W_t| = |(1 - \phi B)^{-1} \epsilon_t| - |(1 - \phi B)^{-1} \eta_t|.
\]

Multiplying both sides by \( (1 - \phi B) \) and setting \( Y_t = |X_t| - |W_t| \) and \( \epsilon_t = |\epsilon_t| - |\eta_t| \) yields

\[
(1 - \phi B)Y_t = \epsilon_t,
\]

an AR(1) process.

Under SE loss, it can be shown that the loss differential is an AR(2) process.\(^2\) Again, using the series described in Eq. (A1), the loss differential under SE loss becomes

\[
[(1 - \phi B)X_t^2] - [(1 - \phi B)W_t^2] = \epsilon_t^2 - \eta_t^2,
\]

which becomes

\[
(1 - \phi B)^2(X_t^2 - W_t^2) = (1 - 2\phi B + \phi^2 B^2)(X_t^2 - W_t^2)
= (1 - \alpha_1 B - \alpha_2 B^2)(X_t^2 - W_t^2)
= \epsilon_t^2 - \eta_t^2,
\]

which is an AR(2) process with autocorrelation coefficients \( \alpha_1 = 2\phi \) and \( \alpha_2 = \phi^2 \).

A more general result for the difference between two ARMA series can be made following a similar path as the examples above, using the difference operator notation. Suppose \( X_t \) follows an ARMA(\( r, m \)) and \( W_t \) an ARMA(\( s, n \)) model, each with different autoregressive and moving average parameters. Then,

\[
(1 - \phi_1^r B - \cdots - \phi_r^r B^r)X_t = (1 - \theta_1^r B - \cdots - \theta_m^r B^m)\epsilon_t, \tag{A3}
\]

and

\[
(1 - \phi_1^s B - \cdots - \phi_s^s B^s)W_t = (1 - \theta_1^s B - \cdots - \theta_n^s B^n)\eta_t, \tag{A4}
\]

where \( \epsilon_t \) and \( \eta_t \) are the usual white-noise processes. Multiplying Eq. (A3) by \( (1 - \phi_1^r B - \cdots - \phi_r^r B^r) \) and Eq. (A4) by \( (1 - \phi_1^s B - \cdots - \phi_s^s B^s) \) and subtracting results in

\(^2\) If the autocorrelation coefficients are different, then the loss differential series would be an ARMA(4, 2) process.
The left-hand side of Eq. (A5) results in an order \( p = r + s \) polynomial for the difference operators, so that it can be written in the form \((1 - \phi_1 B - \cdots - \phi_r B^r)(1 - \phi_1 B - \cdots - \phi_s B^s)Y_t\). The difference operators on the right-hand side can be rewritten as a single polynomial with order equal to the greater of \( s + m \) and \( r + n \) [i.e., \((1 - \theta_1 B - \cdots - \theta_m B^m)B^{n+|r+n|})\nu_t\)]. Thus, the series \( Y_t = X_t - W_t \) follows an ARMA\((p, q)\) model, where \( p = r + s \) and \( q = \max\{s + m, r + n\} \).


