

A new approach to testing forecast predictive accuracy

Eric Gilleland* and Gregory Roux

Research Applications Laboratory, National Center for Atmospheric Research, Boulder, CO, USA

ABSTRACT: The Diebold–Mariano test for predictive accuracy has been used widely and adapted for economic forecasts, but has not seen much activity in weather forecast verification. The technique is applied to both simulated verification sets as well as weather data at eight stations in Utah, and a loss function based on dynamic time warping (DTW) is used. Results of the simulation experiment show that the DTW technique can be useful if timing errors are the concern. Real test cases demonstrate the difficulty in automating some of the more advanced methods proposed here, but also show the utility in even the most basic test, which is an improvement over similar tests that do not account for temporal and/or contemporaneous correlation.

KEY WORDS forecast verification; predictive accuracy; dynamic time warping; Diebold–Mariano test; hypothesis testing; statistical inference; time series

Received 1 April 2014; Revised 15 October 2014; Accepted 17 October 2014

1. Introduction

Statistical summaries are sought out in forecast verification in order to inform about forecast performance. Often, these studies are reported without any consideration of sampling uncertainty, which is necessary for making informed decisions about a forecast's accuracy (see, e.g. Jolliffe, 2007). Hamill (1999), Jolliffe (2007) and Gilleland (2010) provide useful summaries to many of the common techniques for accounting for sampling uncertainty. When sampling uncertainty is taken into consideration, such techniques often require temporal independence assumptions that typically are not met by meteorological variables. Block bootstrapping (e.g. Wilks, 1997; Gilleland, 2010) and variance inflation (e.g. Wilks, 1997) are two common, and relatively simple, approaches that account for temporal dependence. The Diebold–Mariano (DM) test (Diebold and Mariano, 1995), and modifications thereof, has been extensively used in the field of economic forecasting (e.g. Christoffersen, 1998; Diebold *et al.*, 1998; Lettau and Ludvigson, 2001; Poon and Granger, 2003). The test compares the accuracy of two competing forecasts and directly accounts for temporal dependence with few assumptions. As far as is known, this test has not been applied in a weather forecasting setting, apart from Hering and Genton (2011) and Gilleland (2013). In both of these studies, the emphasis is on analysing forecasts spatially. Here, only the univariate time series setting is investigated.

Timing errors are another type of error that are often not considered when verifying forecasts quantitatively. Traditional forecast verification is conducted for forecast and verification pairs at the same time points (e.g. Hamill, 1999), but it can be the case that a forecast could tend to be too fast or too slow. Incorporation of such timing errors could provide valuable information on forecast performance.

The present study demonstrates the DM test in conjunction with a technique that assesses timing errors in the forecast in addition to intensity errors *via* dynamic time warping (DTW). DTW has been used extensively in fields such as speech processing (e.g. Rabiner *et al.*, 1978; Sakoe and Chiba, 1978; Deller *et al.*, 1999), data mining (e.g. Berndt and Clifford, 1994; Keogh and Ratanamahatana, 2004), bioinformatics (e.g. Aach and Church, 2001), and many others. In the present context, it is used as a 1D analogue to the various field morphing techniques applied to date to spatial forecast verification (e.g. Hoffman *et al.*, 1995; Hoffman and Grassotti, 1996; Alexander *et al.*, 1998, 1999; Nehrkorn *et al.*, 2003; Keil and Craig, 2007, 2009; Gilleland *et al.*, 2010a, 2010b; Marzban and Sandgathe, 2010); in concert with the prediction comparison test applied here, the entire process is a 1D analogue to the spatial prediction comparison test using image warping loss introduced by Gilleland (2013).

In the 1D setting, the only source that uses DTW for weather forecast verification is that of Lin *et al.* (2010) who assess the error in air quality models where special concern is on identifying the point source for a pollutant's release.

2. Methods

In Section 2.1, the DM test procedure is described, as well as a proposed modification for the test. This background is followed in Section 2.2 with a description of the proposed DTW loss function.

2.1. The Diebold–Mariano test

Diebold and Mariano (1995) introduced a statistical test for the null hypothesis of equal forecast accuracy between two competing models (hereafter, the DM test). Their test has been used extensively in the literature (*cf* Hering and Genton, 2011). The DM test can be used with any loss function, such as straight differences, absolute differences, squared differences, and so on, as well as skill functions, such as correlation. Moreover, the test makes no distributional assumptions on the forecast errors, and

* Correspondence: E. Gilleland, Research Applications Laboratory, National Center for Atmospheric Research, Boulder, CO, USA. E-mail: ericG@ucar.edu

incorporates temporal autocorrelations, as well as any correlation between the two series. Numerous adaptations to the test have been made (e.g. West, 1996; Harvey *et al.*, 1997; Dell'Aquila and Ronchetti, 2004; McCracken, 2004; Giacomini and White, 2006; Hering and Genton, 2011).

The test is conducted as follows. Let $\hat{x}_{11}, \dots, \hat{x}_{1n}$ and $\hat{x}_{21}, \dots, \hat{x}_{2n}$ be competing forecasts for a variable, such as temperature valid at time t , which is denoted as x_1, \dots, x_n . Finally, let $g(x_t, \hat{x}_{it})$, $i = 1, 2$ be a loss function that measures the accuracy or skill of the forecast (e.g. if g is simple loss, then $g(x_t, \hat{x}_{it}) = \hat{x}_{it} - x_t$). The null hypothesis of equal forecast accuracy for two sets of forecasts is then simply that the two loss functions are expected to be the same on average. Technically, the hypothesis is written as:

$$H_0 : E[g(x_t, \hat{x}_{1t})] = E[g(x_t, \hat{x}_{2t})] \quad (1)$$

where E denotes the expected value over all time points. To test whether the accuracy of the two competing forecast models is equal, the difference in loss functions, $d_t = g(x_t, \hat{x}_{1t}) - g(x_t, \hat{x}_{2t})$, is used (so, if forecast 1 is better, then the loss differential will be negative on average, and if forecast 2 is better, the loss differential will be positive on average). The hypothesis in Equation (1), then, is equivalent to:

$$H_0 : E[d_t] = 0 \quad (2)$$

The difference in loss functions, d_t , is referred to as the loss differential and it is assumed to be covariance stationary (i.e. for a given lag of time, the covariance of the time series between any two values separated by the same lag is the same), and it has short memory (i.e. the covariance of two values is zero beyond a certain lag). Subsequently, the asymptotic distribution of the sample mean loss differential, $\bar{d} = \frac{1}{n} \sum_{t=1}^n d_t$, as $n \rightarrow \infty$ is $N(\mu_d, 2\pi s_d(0))$, with μ_d the population mean loss differential and $s_d(0)$ the spectral density of the loss differential at frequency zero.

In order to perform the test in Equation (2), a sample test statistic is sought, such that it can be assumed to follow a particular distribution. This distributional assumption is for the test statistic, and not for the underlying variables. For this problem, the large-sample test statistic for forecast accuracy is given by:

$$S = \frac{\bar{d}}{\sqrt{2\pi \hat{s}_d(0)/n}} \quad (3)$$

where $\hat{s}_d(0)$ is a consistent estimator of $s_d(0)$. The estimator, $\hat{s}_d(0)$, is found by a weighted sum of the available sample autocovariances for a k -step forecast. Hering and Genton (2011) modify this estimator in order to avoid problems with estimating negative values for $s_d(0)$. Specifically, they suggest fitting a parametric covariance model to the empirical autocovariances, which is guaranteed to be positive definite, and summing over the lags of the model instead of the empirical autocovariances. Using \hat{C} to denote the estimated covariance function, the estimate becomes:

$$2\pi \hat{s}_d^p(0) = \hat{C}(0) + 2 \sum_{\tau=1}^{n-1} \hat{C}(\tau) \quad (4)$$

where τ is the temporal lag so that the estimated covariance is between the value at time zero and that at time τ . Equation (4) yields a parametrically estimated test statistic replacing $\hat{s}_d(0)$ with $\hat{s}_d^p(0)$ in Equation (3). Hering and Genton (2011) propose using an exponential covariance model of the form $C(\tau) = \sigma^2 \exp(-\tau/\theta)$, where σ^2 is the marginal variance, and the practical range of dependence (i.e. the range beyond

which the correlation is <0.05) is governed by the parameter θ . Both of these parameters, σ and θ , need to be estimated from the observed loss differential series, d_t . To this end, nonlinear least squares estimation is employed, which requires numerical optimization to find simultaneously the desired estimates (see Bates and Watts, 1988; Bates and Chambers, 1992; for more details about nonlinear least squares estimation). Because the DM test is valid for any loss function, in addition to investigating some of the usual loss functions (absolute error, square error, and correlation skill), it is interesting to also employ a loss function that accounts for timing errors in the forecast. In this vein, the technique of DTW is employed here.

2.2. DTW loss function

DTW was first introduced by Bellman and Kalaba (1959) and is a method for finding the optimal alignment between two time series. Given two time series as before, where $\hat{x}_1, \dots, \hat{x}_n$ is the test series and x_1, \dots, x_n the reference series, an $n \times m$ matrix M is constructed such that the entries in the i^{th} row and j^{th} column represent the distance between x_i and \hat{x}_j . The DTW is the path through M that minimizes the warping cost, and satisfies the following constraints (cf Giorgino, 2009; Lin *et al.*, 2010). Let w_1, \dots, w_k represent the warping path, then the conditions are:

1. boundary conditions: the warping path must start at the first row and first column, $w_1 = (1, 1)$, and end in the last row and last column, $w_k = (m, n)$;
2. continuity: the difference in distance between successive steps in the path must be less than or equal to one in order to confine allowable steps in the warping path to be neighbouring points (i.e. if $w_i = (a, b)$ and $w_{i-1} = (a', b')$, then $a - a' \leq 1$ and $b - b' \leq 1$);
3. monotonicity: further ensure that $a \geq a'$ and $b \geq b'$ in order that the path be monotonically ordered with respect to time, thereby avoiding meaningless loops.

The first condition can be relaxed in order to compute partial time series matches (cf Giorgino, 2009, Section 3.5). The warping cost can be computed relatively quickly using dynamic programming (Myers *et al.*, 1980; Giorgino, 2009), and is an accumulated distortion between the test and the reference series. The distance between points can be given by any of a number of functions, although the Euclidean distance is the most common choice (Giorgino, 2009).

The final result of the DTW is a mapping of time indices between the test and reference series describing a better alignment. For example, if a forecast tends to be too early or too late in predicting an event, then after warping, such predictions should be precisely on time. DTW is a 1D analogue to the more complicated image warping techniques (e.g. Aberg *et al.*, 2005; Gilleland *et al.*, 2010a, 2010b; Gilleland, 2013), and can be used to obtain warps in more than one dimension, thereby enabling the possibility of obtaining spatial warps (e.g. Lin *et al.*, 2010).

The DTW loss function is then given by:

$$g(x_t, \hat{x}_{it}) = f(t, w_{it}(t)) + h(x_t, \hat{x}_{w_{it}(t)}) \quad (5)$$

where the first term, f , is a loss function (e.g. absolute error and Euclidean distance) on the temporal movements and h is a loss function on the two processes after re-alignment.

The test for predictive accuracy described here with DTW loss is a 1D analogue to the spatial prediction comparison test with image warping loss as introduced in Gilleland (2013).

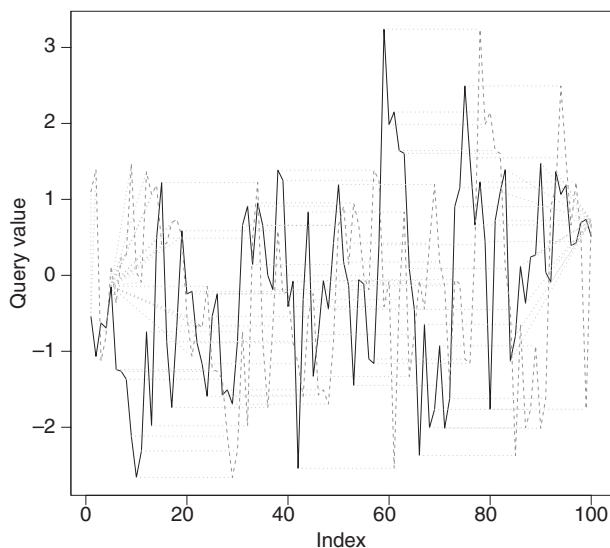


Figure 1. Example of DTW using a simulated AR(3) process (simulated according to Equation (6), solid black line) and the same process shifted in time (dashed grey line). The dotted grey lines indicate point-to-point movements from the warping.

Although the DM test procedure has been well studied in other fields, it is not known if it is being used in the arena of weather forecast verification (outside of Hering and Genton (2011) who employ their spatial extension of the test to wind energy data, and Gilleland (2013) who uses the spatial version for quantitative precipitation forecasts). Further, it is not known if any application combines the DTW framework with the DM test in order to account for timing errors.

Analyses are performed using the dtw (Giorgino, 2009) and verification (NCAR - Research Applications Laboratory, 2012) packages in R (R Core Team, 2012) (see Appendix).

Figure 1 shows a simulation example of DTW. Data are simulated from an autoregression model with three lag terms (i.e. an AR(3) model), with lag co-efficients equal to 0.8, -0.2 and 0.1. That is, the simulated time series, $\{z_t\}_{t=1}^n$, with $n = 100$ is drawn from the model:

$$\begin{aligned} z_t &= \rho_1 z_{t-1} + \rho_2 z_{t-2} + \rho_3 z_{t-3} + \varepsilon_t \\ &= 0.8z_{t-1} - 0.2z_{t-2} + 0.1z_{t-3} + \varepsilon_t \end{aligned} \quad (6)$$

where $\varepsilon_t \sim N(0,1)$. A second simulated time series is made by simply shifting $\{z_t\}_{t=1}^n$ 20 places to the left. The beginning of the series $\{z_t\}_{t=1}^n$ is tacked on to the end of the shifted series. It can be seen that for the bulk of the data, a simple shift by 20 time points is made, and a slightly more complicated displacement is invoked for the beginning and end, which differ by more than a mere temporal shift.

3. Data

The testing procedure is applied to both simulated and real forecast data sets. The real data set is based on forecasts with lead time of 48 h for eight observing sites in the US state of Utah (Liu *et al.*, 2009). Thirty different forecasts exist for each of the sites, and could be considered as a low-grade ensemble for each location, though they will be treated individually here. Model cycles begin at 0000 UTC, 0600 UTC, 1200 UTC and 1800 UTC, but only 0000 UTC is considered here. Variables

considered include: temperature, u- and v-winds, pressure and relative humidity.

An example of data and model output for one station is shown in Figure 2. Grey lines in the figure indicate that the values were missing and have been interpolated. Temperature and relative humidity were both interpolated by fitting a harmonic regression to the data to predict the values where they were missing, and then a small amount of random noise was added. For u- and v-winds, as well as pressure, an AR(3) model (Equation (6), but where the correlation co-efficients, ρ_k , $k = 1-3$, are estimated from the data) is used to predict where the values are missing. Again, small random noise is subsequently added to the predictions. In most cases, the imputations appear reasonable. In practice, it may be best to simply apply the test to the bulk of data where no missing values occur, or take greater care in their imputation. However, the investigation of the verification method as opposed to the determination of the best ensemble member is the primary concern; so all the values here are interpolated for simplicity.

Simulations are primarily used to demonstrate the potential utility of applying a loss function that incorporates information from a DTW. Hering and Genton (2011) investigated the size and power of their modification to the DM test by simulating the errors themselves because otherwise one would need to build a prediction model, of which an infinite number could be considered. This study does not attempt to infer about the size or power of the test with the DTW loss function, but a similar type of analysis is performed in order to obtain information about how the test treats different types of timing and intensity error combinations in a manner more similar to that of Ahijevych *et al.* (2009), except that shifting and additive errors imposed on the forecast are not merely perturbations of the verification series, but are instead perturbations of a separate, highly correlated, series. With this background, data are simulated in the following manner:

- 1 Three sets of AR(1) models of size n are simulated using `arima.sim` from package `stats` (R Core Team, 2012). The first is to be the ‘verification’ simulation, z , the second is forecast 1, \hat{z}_1 , and the third is forecast 2, \hat{z}_2 .
- 2 \hat{z}_1 and \hat{z}_2 are correlated with z using a correlation of 0.65.
- 3 \hat{z}_1 and \hat{z}_2 are correlated with each other using a more modest correlation of 0.2.
- 4 \hat{z}_2 is shifted k units to the left and an additive intensity error is applied.

Step 1 involves simulations from the model $z_t = \rho z_{t-1} + \varepsilon_t$, where $\varepsilon_t \sim N(0,1)$ and ρ is varied to be 0.2, 0.4, 0.6 and 0.8 going from a weak temporal dependence to a strong one. The sample size increases from 10 to 100 by 10 (i.e. 10, 20, 30, ..., 100).

Steps 2 and 3 are carried out by column binding the two pertinent simulations together to form a $n \times 2$ matrix \mathbf{Z} , transposing and then right multiplying by \mathbf{R}^T . The two resulting columns of the transposed product are now correlated random vectors. More precisely, let:

$$\mathbf{R}^2 = \begin{bmatrix} 1 & \varphi \\ \varphi & 1 \end{bmatrix} \quad (7)$$

where in the case of step 2, $\varphi = 0.65$, and for step 3, $\varphi = 0.2$. The two correlated random vectors are obtained from the columns of $(\mathbf{R}^T \mathbf{Z}^T)^T$.

Step 4 is a simple shift by 0, 2, 10 and 20 places to the left (for brevity, results for a shift of 2 are not shown), and step 5 is performed by adding $a\theta$ to the result where a is varied over 0, 1, 5 and 10, and θ is a random vector of length n sampled from

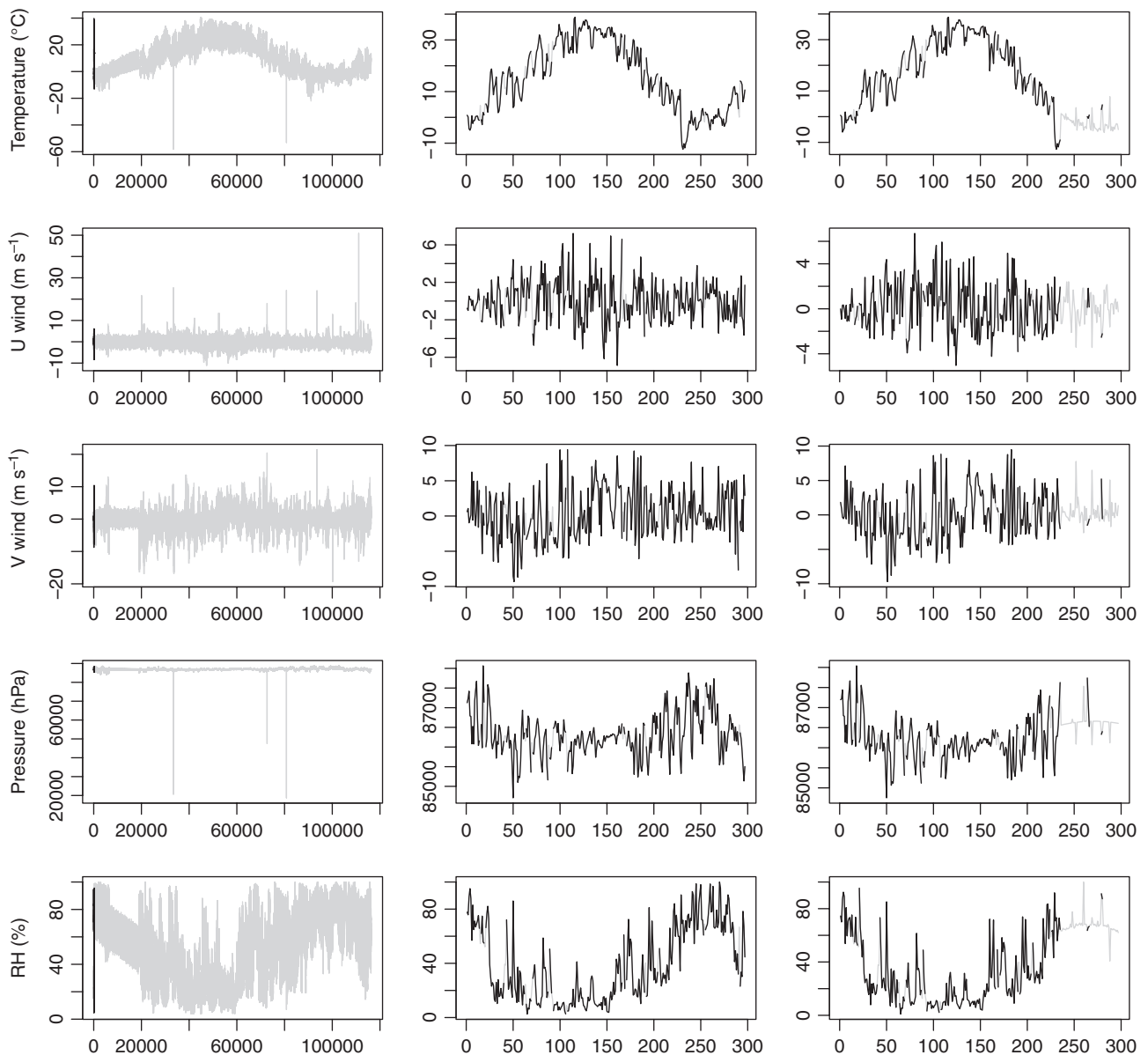


Figure 2. Observations (first column) and 0 h model output (GFS MR220, middle column and NAM MCKF2, third column) from station 1 in Utah. Grey lines indicate where data/model output were missing and have been interpolated.

a standard normal distribution function. In particular, when the shift is 0, no shifting is conducted. When $a = 0$, no additive error is applied. If $a > 0$, then the variance of the random additive errors is given by a^2 .

Figure 3 shows an example of such simulated data using a sample size of 100. The first two rows show sample autocorrelation function (ACF) and partial autocorrelation function (PACF) plots, which diagnose properties about the temporal dependence structure of a time series. For more details about these diagnostic plots, see Brockwell and Davis (1996) or Gilleland (2010) for an accessible introduction. Briefly, for each time lag, an ACF plot is a plot of the sample ACF, which is the estimated correlation among pairs of points separated by the same time lag, plotted against increasing time lags starting from zero; the ACF at lag zero will always be one. If a sample of size 100 (as in the figure), were independent in time, then one would expect five values to fall outside of the 95% confidence bounds (dashed lines). The PACF is similar in spirit to the ACF plot, but plots the

correlation between prediction errors for values at lag k against those at lag zero, and again plotted against increasing lags beginning with one. For an $AR(p)$ model, the PACF is expected to be zero for lags larger than p .

In each case, strong correlations exist between each simulated series. However, when \hat{z}_2 is shifted by 20 spaces and has error added to it with $a = 5$, (labelled \hat{z}_3 in the figure), the resulting correlation is nearly zero.

4. Results

Results from the simulation experiment employing DTW with absolute error loss are displayed in Figures 4 and 5. Figure 4 shows the percentage of rejected null hypotheses from the testing procedure applied to the simulated series. The top row is for cases A–L, and the bottom row shows cases M–X as described in Table 1. The left column is for significance levels of 0.01 and the

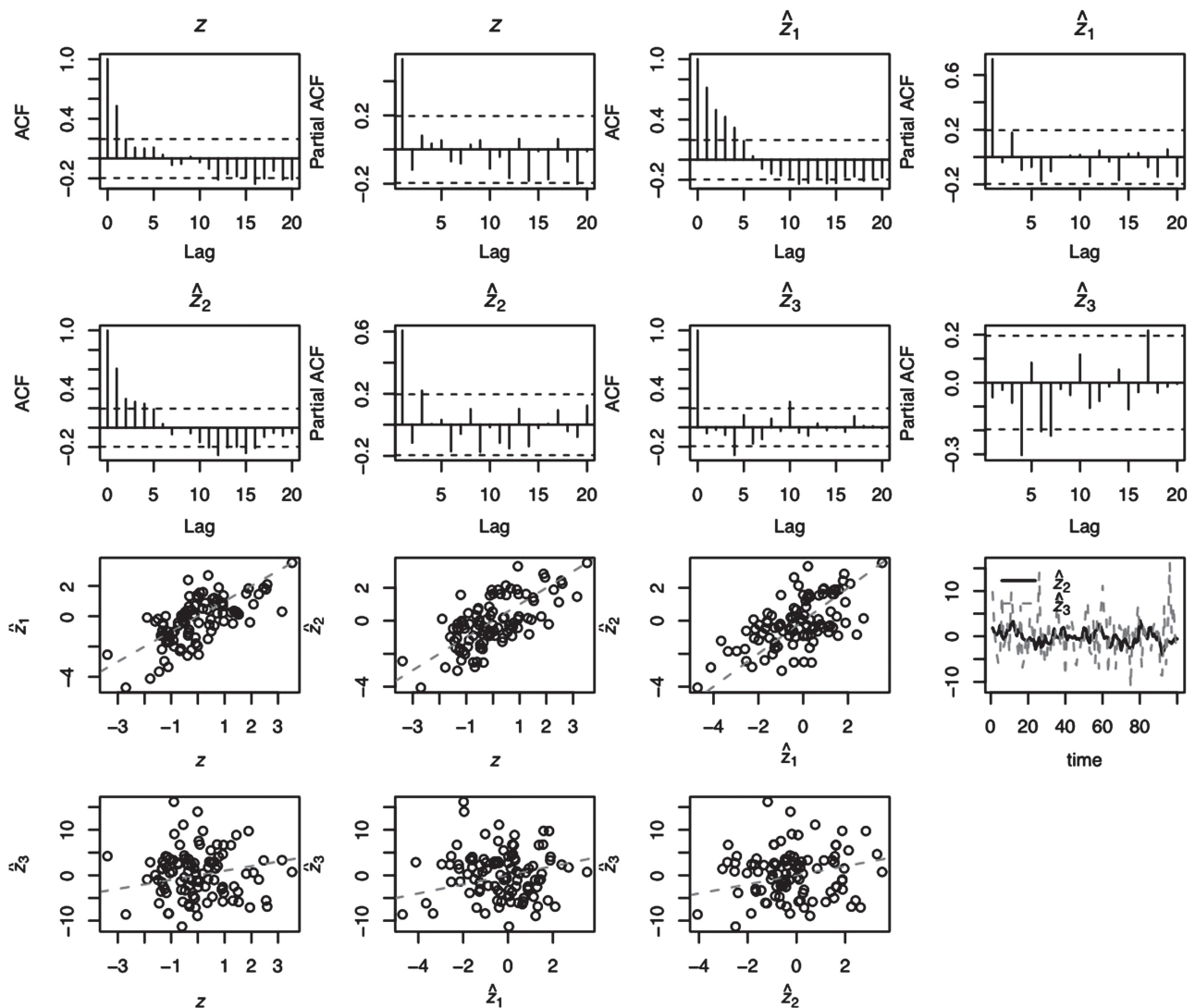


Figure 3. Example of a simulated AR(1) series using a correlation co-efficient of 0.6 and sample size of 100. Grey dashed lines on scatter plots are the one-to-one lines. The simulation of \hat{z}_2 is also shifted 20 spaces to the left and an additive error is applied with $a = 5$; the resulting series is denoted with a subscript 3 here. For this example, the empirical correlation between z and \hat{z}_1 , as well as between z and \hat{z}_2 is about 0.66. Between \hat{z}_1 and \hat{z}_2 the correlation is about 0.59. The resulting empirical correlation of z and \hat{z}_3 is nearly zero.

right for levels of 0.05. It is not important to discern the individual cases in this figure, but rather to note the general rejection rates.

As might be expected from inspection of the example in Figure 3, the null hypothesis is rejected more often for cases where one of the simulated forecasts is shifted. Although similar for additive errors, the results curiously do not suggest that the addition of additive errors implies more rejected null hypotheses. Although this simulation study does not inform about the size of the test because it is not clear which simulations should result in a rejection of the null hypothesis and which should not, a horizontal line through $1 - \alpha$, where α is the significance level of the test, is given as a guide. That is, if one model was superior to another in terms of this test, then the null hypothesis should be rejected 100% of the time. However, because of sampling error, some of the tests are expected to be in error. This error is controlled to occur at the rate α . So, if models could be simulated such that one is known to be better, rejections should be in the order of $(1 - \alpha) \cdot 100\%$ (i.e. the lines in the figure would be very close to the grey dashed lines). Because it is not clear how to simulate such series, these dashed lines are

included only to gain a rough idea of how the procedure works in general. On the whole, it is expected that \hat{z}_2 is generally the worse forecast, so a high null hypothesis rejection rate is desired and acquired from these simulations. As expected, simulations with low or no shifting and low or no additive errors yield fewer rejections.

Inspection of Figure 5, which shows histograms of the test statistics, reveals that more often than not, simulated forecast \hat{z}_1 is the better of the two forecasts (because $\bar{d} < 0$), which is reasonable given that \hat{z}_2 is simulated to be less correlated (at least, after shifting and adding additional error) with the verification series. Cases A, B, M and N show more symmetry around zero, which is also reasonable, as these cases do not involve any shifting or adding of additional error.

The DM statistics (Equation (3)) for the real verification set (cf Figure 2) are shown in Figure 6 (for temperature variable at station 1 valid at 0000 UTC) for every combination of stations as a symmetrical image matrix. Member identifiers for members labelled 1–30 are given in Table 2 (see Liu *et al.*, 2009 for more details about the forecast members). Only the upper

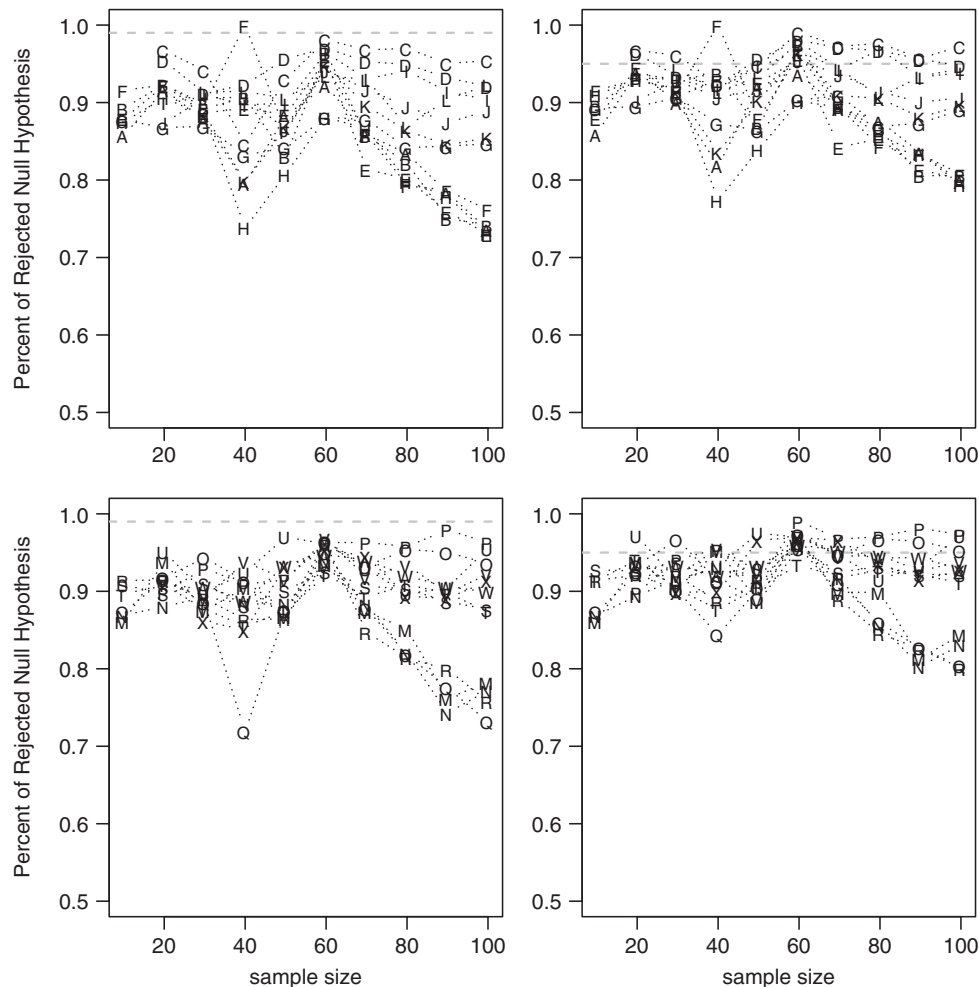


Figure 4. Percent of rejected null hypotheses (using absolute error with DTW loss) for significance levels of 0.01 (first column) and 0.05 (second column) for the different letter-labelled cases described in Table 1: AR(1) correlation co-efficient, shift and amplitude amount combinations analysed with letter codes. Combinations are not meant to be exhaustive, but represent a relatively wide variety of case types. M–X are the same as A–L except for the AR(1) correlation co-efficients (top row are cases A–L and bottom row are cases M–X). Grey dashed lines indicate the significance level for the test.

triangle is plotted because the lower triangle would give redundant information. Member 19 is missing for this station and valid time. Loss differentials are constructed so that the loss function for the higher-numbered member is subtracted from that of the lower-numbered member. In this way, negative values imply superior performance of the lower-numbered member, whereas positive values imply that the higher-numbered member is superior according to the statistic.

Investigating the colours vertically for member 2 in Figure 6(a) (see vertical dashed line), it is clear that other members generally perform worse in terms of the DM test statistic under absolute error loss. The worst performing member for the temperature variable (station 1 and valid time 0000 UTC), in general, appears to be member 13. Figure 6(a) does not account for statistical significance. Even at the 5% level, not many of the results are statistically significant. Figure 6(b) shows the same results, but only those statistics that are significant at the 5% level are displayed; most of the member 13 results are also significant.

Figure 7 shows similar results, but for u-wind. The majority of values are positive indicating that the models on the ordinate axis are generally better in terms of the DM statistic, many of which are also statistically significant.

Table 1. AR(1) correlation co-efficient, shift amount and amplitude amount combinations analysed with letter codes. Combinations are not meant to be exhaustive, but represent a relatively wide variety of case types. M–X are the same as A–L except for the AR(1) correlation co-efficients.

A	$\rho = 0.6$; shift = 0; $a = 0$	M	$\rho = 0.2$; shift = 0; $a = 0$
B	$\rho = 0.8$; shift = 0; $a = 0$	N	$\rho = 0.4$; shift = 0; $a = 0$
C	$\rho = 0.6$; shift = 10; $a = 0$	O	$\rho = 0.2$; shift = 10; $a = 0$
D	$\rho = 0.8$; shift = 10; $a = 0$	P	$\rho = 0.4$; shift = 10; $a = 0$
E	$\rho = 0.6$; shift = 10; $a = 0$	Q	$\rho = 0.2$; shift = 10; $a = 0$
F	$\rho = 0.8$; shift = 0; $a = 1$	R	$\rho = 0.4$; shift = 0; $a = 1$
G	$\rho = 0.6$; shift = 0; $a = 5$	S	$\rho = 0.2$; shift = 0; $a = 5$
H	$\rho = 0.8$; shift = 0; $a = 5$	T	$\rho = 0.4$; shift = 0; $a = 5$
I	$\rho = 0.6$; shift = 10; $a = 5$	U	$\rho = 0.2$; shift = 10; $a = 5$
J	$\rho = 0.8$; shift = 10; $a = 5$	V	$\rho = 0.4$; shift = 10; $a = 5$
K	$\rho = 0.6$; shift = 20; $a = 5$	W	$\rho = 0.2$; shift = 20; $a = 5$
L	$\rho = 0.6$; shift = 20; $a = 0$	X	$\rho = 0.2$; shift = 20; $a = 0$

Results are similar for the Hering and Genton (2011) modification of the DM test (not shown), except for fewer significant results, which may, in part, be an issue pertaining to difficulties

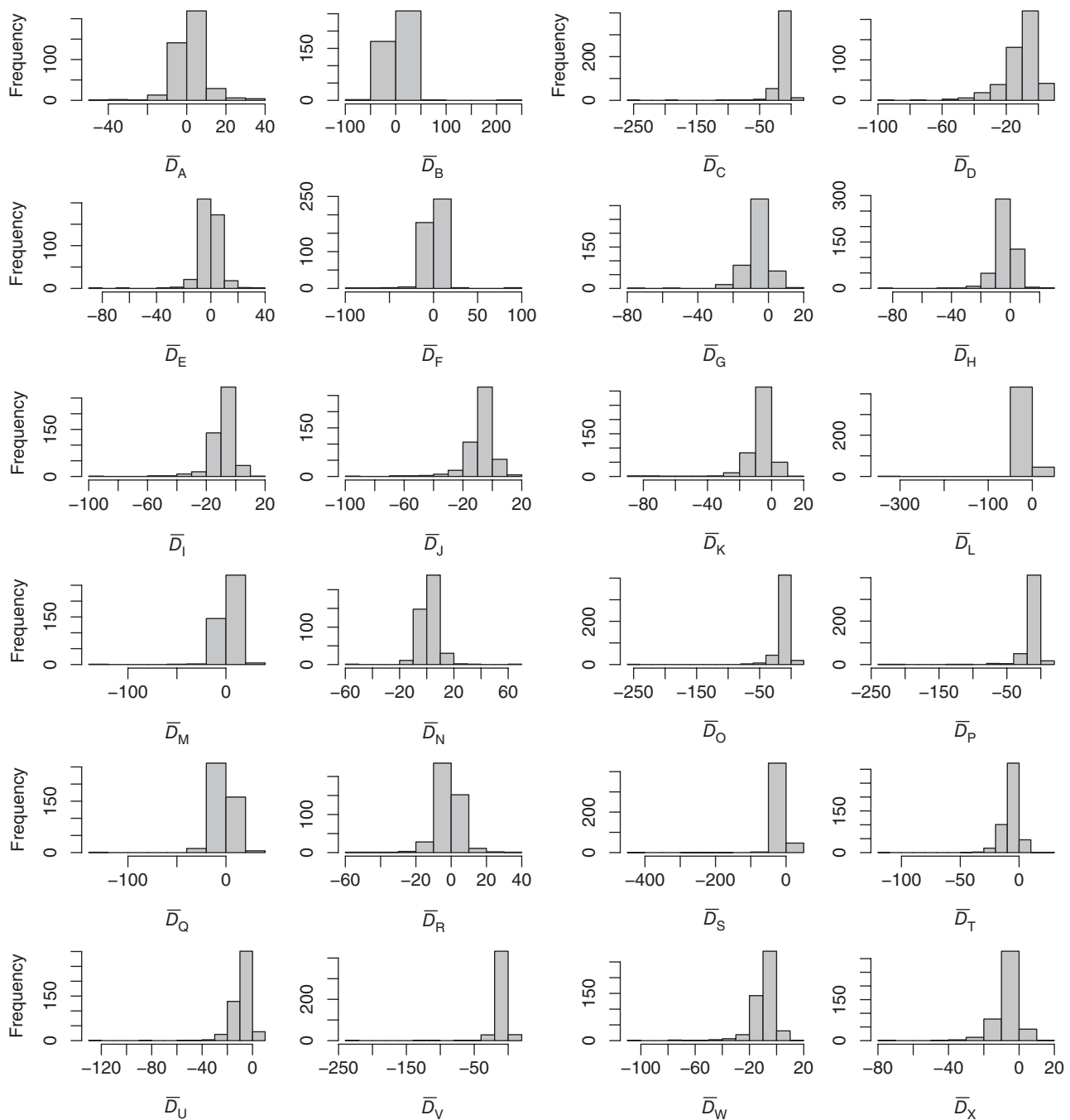


Figure 5. Histogram of the test statistics (using absolute error with DTW loss) for the simulated data sets of size 100. Subscripts in abscissa labels represent the specific simulation from Table 1: AR(1) correlation co-efficient, shift and amplitude amount combinations analysed with letter codes. Combinations are not meant to be exhaustive, but represent a relatively wide variety of case types. M–X are the same as A–L except for the AR(1) correlation co-efficients. Negative values imply that \hat{z}_2 is the worse of the two forecasts.

in automatically fitting a parametric autocovariance model to the empirical autocovariances.

Similarly, applying the DTW with absolute error loss (also not shown), no significant results are found (at the 5% level), which suggests that, according to this test, the members have similar temporal displacement errors so that one member is not statistically significantly superior to another in terms of timing.

Figure 8 shows results across all stations and valid times for the pressure variable where one model was significantly better than the corresponding model according to the Hering and Genton (2011) modification to the DM test using DTW with absolute

error loss as in Equation (5). Figure 8(a) indicates the number of times the member on the ordinate axis significantly (5% level) outperforms the corresponding member on the abscissa, and Figure 8(b) the number of times the member on the abscissa outperforms that on the ordinate. For example, the figure shows that member 5 is never outperformed, according to this test at the 5% level, by any other member for the pressure variable, but it statistically significantly outperforms most other members on the order of 30–40 times (although, this is not very many times considering it is across all eight stations with 49 valid times, meaning that it outperforms the other members about 7–10%

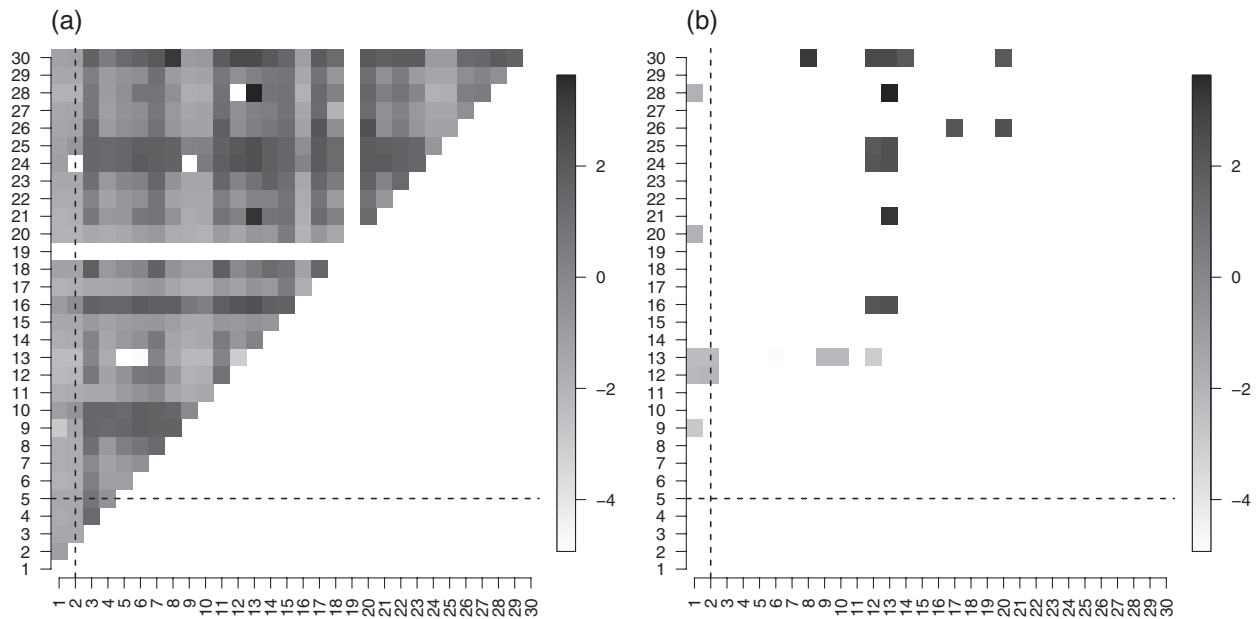


Figure 6. Diebold–Mariano (Equation (3)) statistics (using absolute error loss) comparing every (paired) combination of the 30 ensemble members for temperature valid at 0000 UTC for station 1. (a) shows all values of the statistic; (b) shows just those that are statistically significant at the 5% level. Where the dashed lines cross, indicates the two models depicted in Figure 2. The higher numbered member's loss is subtracted from the lower numbered members' loss in the calculation of the DM statistic so that lower (higher) values imply that the associated model on the abscissa (ordinate) is superior to the corresponding model on the ordinate (abscissa) axis as the lower (redundant) triangle is not plotted.

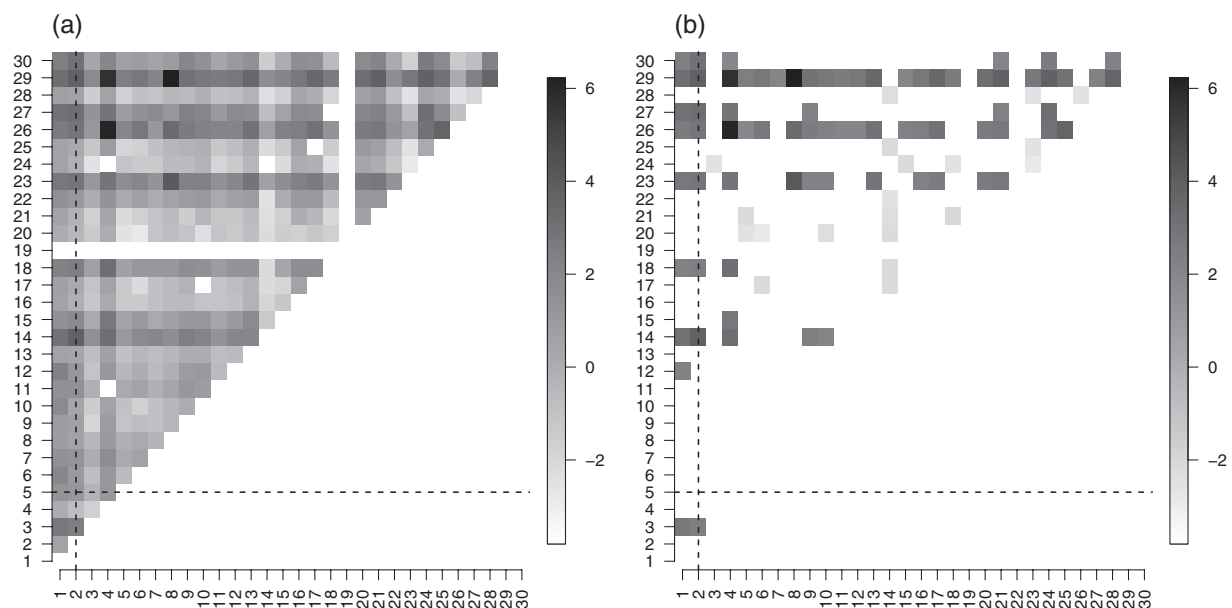


Figure 7. Same as Figure 6, but for u-wind.

of the time). Member 6 has similar results, except that it is outperformed at some times by member 28; members 20 and 21 also both fare very well.

For the other variables, few significant results (at the 5% level) are found using DTW with absolute error loss; none for temperature, 22 for u-wind, three for v-wind and only one for relative humidity. This does not imply that application of DTW to the series does not improve the verification results for individual members. The reduction in error for temperature, in terms of mean square error, ranges from about 40% to ~84% across all stations, models and forecast times. For u-wind, the improvements are greatest at about 49–89%. Relative humidity

has a minimum improvement on the order of 22%, and v-wind about 31%, whereas both have as high as 88% improvement. Such large improvements in mean square error after applying DTW indicate that timing errors may be a concern for these members.

5. Conclusions

A useful method is introduced for testing competing forecast time series against the same observation that originates from economics and statistics literature (primarily from Diebold and

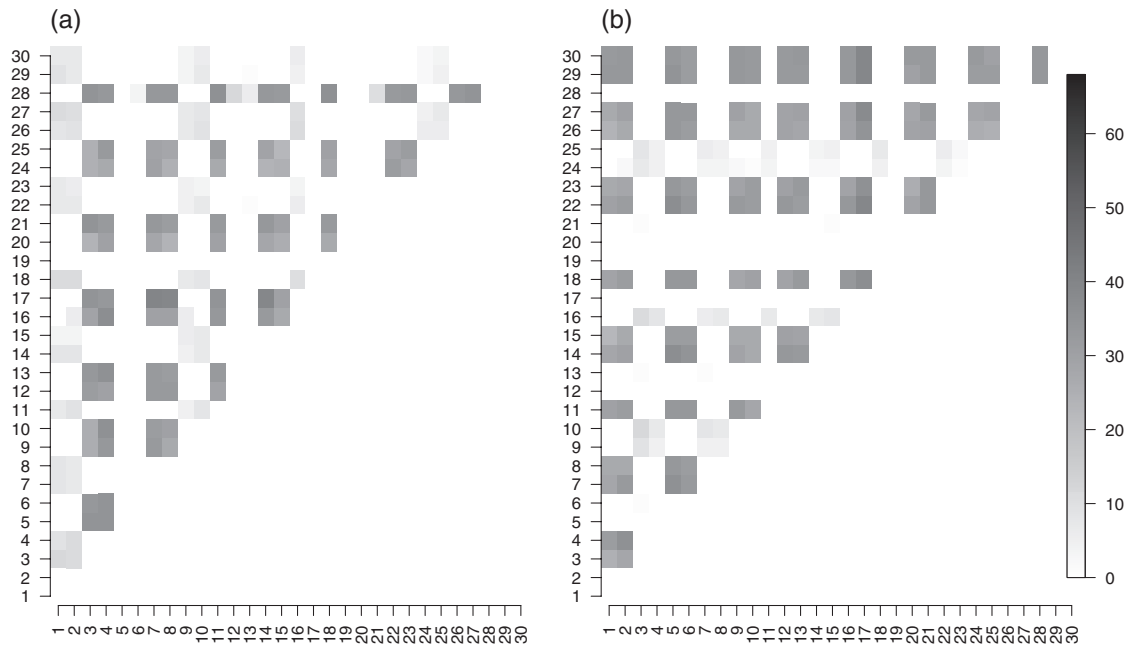


Figure 8. Number of times a member is significantly better than another according to DTW loss with absolute error for pressure (all stations, all forecast times). (a) The number of times that the members on the ordinate axis were significantly (5% level) better than those on the abscissa. (b) The number of times the members along the abscissa were significantly better than those on the ordinate. White spaces indicate a test that did not yield a result (e.g. because a parametric model or DTW could not be easily fit); the lower right triangle would have been redundant, and so is also left white.

Table 2. Ensemble member identifiers.

Model number	Model ID	Model number	Model ID
1	GFS_MCBM1	15	NAM_WPRUC
2	GFS_MR220	16	GFS_MLMLM
3	GFS_WCBMJ	17	GFS_MUPRA
4	GFS_WPBOU	18	GFS_WCTRL
5	NAM_MCKF2	19	GFS_WRCAM
6	NAM_MMGOB	20	NAM_MLMLP
7	NAM_WCTRL	21	NAM_MRCLD
8	NAM_WPQNS	22	NAM_WMWS6
9	GFS_MCTRL	23	NAM_WRGOD
10	GFS_MRCCM	24	GFS_MLPLP
11	GFS_WPMYN	25	GFS_MUPRB
12	NAM_MCTRL	26	GFS_WMTHO
13	NAM_MPBLE	27	GFS_WRSLP
14	NAM_WMMOR	28	NAM_MLPLM
15	NAM_WPRUC	29	NAM_WCGDE

Mariano, 1995; Giacomini and White, 2006) for weather forecast verification purposes. It is often of interest to apply a statistical test to determine if one forecast’s verification statistic is significantly better (or worse) than a competing forecast model. However, such testing is difficult because of the presence of temporal, as well as contemporaneous, correlation, which is characteristic of most meteorological variables. Hering and Genton (2011) demonstrated that their modification of the test is robust not only to temporal, but also to contemporaneous correlations (i.e. when the two forecast models are correlated with each other). Use of dynamic time warping (DTW) as an additional tool for evaluating forecast performance will be a welcome option. Simulation results demonstrate that, in principle, the technique yields results

that are meaningful (i.e. inform about forecast performance while accounting directly for timing errors).

In practice, fitting parametric models to the autocovariances for the real data sets was found to be difficult to automate, but it was found that generally the Diebold–Mariano (DM) test worked well. Statistically significant results were few with the testing procedure suggesting that the member forecasts are either very similar to each other, or that the duration in time used is too short; the aim of this study is to demonstrate the testing procedures rather than to analyse the model performance. The pressure variable resulted in the most significant differences in model performance when timing errors are considered simultaneously with intensity errors, though they were still relatively few overall.

The presented testing procedures do not require any underlying assumption of a distribution for the raw variables, nor do they require independence in time. The only distributional assumption is on the resulting average loss differential, which assumes a normal distribution under asymptotic arguments. The DTW allows for informing about timing errors, and in conjunction with the prediction comparison tests, allows simultaneous testing of timing and intensity errors when comparing two models (or comparing against a baseline reference model).

Acknowledgement

The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Supporting information

The following material is available as part of the online article:
Appendix S1. Performing the DM test in R.

References

- Aach J, Church G. 2001. Aligning gene expression time series with time warping algorithms. *Bioinformatics* **17**: 495–508.
- Aberg S, Lindgren F, Malmberg A, Holst J, Holst U. 2005. An image warping approach to spatio-temporal modelling. *Environmetrics* **16**: 833–848.
- Ahiyevech D, Gilleland E, Brown BG, Ebert EE. 2009. Application of spatial verification methods to idealized and NWP gridded precipitation forecasts. *Weather Forecast.* **24**(6): 1485–1497.
- Alexander GD, Weinman JA, Karyampudi VM, Olson WS, Lee ACL. 1999. The effect of assimilating rain rates derived from satellites and lightning on forecasts of the 1993 superstorm. *Mon. Weather Rev.* **127**(7): 1433–1457.
- Alexander GD, Weinman JA, Schols JL. 1998. The use of digital warping of microwave integrated water vapor imagery to improve forecasts of marine extratropical cyclones. *Mon. Weather Rev.* **126**(6): 1469–1496.
- Bates DM, Chambers JM. 1992. Nonlinear models. In *Statistical Models* in S, Chambers JM, Hastie TJ (eds) Chapter 10. Wadsworth & Brooks/Cole: Boca Raton, FL; 421–455.
- Bates DM, Watts DG. 1988. *Nonlinear Regression Analysis and Its Applications*. Wiley: New York, NY; 392.
- Bellman R, Kalaba R. 1959. On adaptive control processes. *IRE Trans. Automat. Control* **4**(2): 1–9.
- Berndt D, Clifford J. 1994. Using dynamic time warping to find patterns in time series. In *AAAI-94 Workshop on Knowledge Discovery in Databases*. 229–248. Technical Report WS-94-03. AAAI: Palo Alto, CA.
- Brokwell PJ, Davis RA. 1996. *Introduction to Time Series and Forecasting*. Springer-Verlag: New York, NY; 420.
- Christoffersen PF. 1998. Evaluating interval forecasts. *Int. Econ. Rev.* **39**(4): 841–862.
- Dell'Aquila R, Ronchetti E. 2004. Robust tests of predictive accuracy. *Metron* **62**: 161–184.
- Deller JR, Hansen JH, Proakis JG. 1999. *Discrete-Time Processing of Speech Signals*. Wiley IEEE Press: Piscataway, NJ; 936.
- Diebold FX, Gunther TA, Tay AS. 1998. Evaluating density forecasts with applications to financial risk management. *Int. Econ. Rev.* **39**(4): 863–883.
- Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **13**: 253–263.
- Giacomini R, White H. 2006. Tests of conditional predictive ability. *Econometrica* **74**: 1545–1578.
- Gilleland E. 2010. Confidence intervals for forecast verification. Technical Note TN-479 STR, NCAR: Boulder, CO; 71. <http://www.ral.ucar.edu/staff/ericg/Gilleland2010.pdf> (accessed 1 May 2014).
- Gilleland E. 2013. Testing competing precipitation forecasts accurately and efficiently: the spatial prediction comparison test. *Mon. Weather Rev.* **141**(1): 340–355.
- Gilleland E, Chen L, DePersio M, Do G, Eilertson K, Jin Y, et al. 2010a. Spatial forecast verification: image warping. Technical Note TN-482 STR, NCAR: Boulder, CO; 23.
- Gilleland E, Lindström J, Lindgren F. 2010b. Analyzing the image warp forecast verification method on precipitation fields from the ICP. *Weather Forecast.* **25**(4): 1249–1262.
- Giorgino T. 2009. Computing and visualizing dynamic time warping alignments in R: the dtw package. *J. Stat. Softw.* **31**(7): 1–24.
- Hamill TM. 1999. Hypothesis tests for evaluating numerical precipitation forecasts. *Weather Forecast.* **14**: 155–167.
- Harvey D, Leybourne S, Newbold P. 1997. Testing the equality of prediction mean squared errors. *Int. J. Forecast.* **13**: 281–291.
- Hering AS, Genton MG. 2011. Comparing spatial predictions. *Technometrics* **53**(4): 414–425.
- Hoffman RN, Grassotti C. 1996. A technique for assimilating SSM/I observations of marine atmospheric storms: tests with ECMWF analyses. *J. Appl. Meteorol.* **35**(8): 1177–1188.
- Hoffman RN, Liu Z, Louis J-F, Grassotti C. 1995. Distortion representation of forecast errors. *Mon. Weather Rev.* **123**(9): 2758–2770.
- Jolliffe IT. 2007. Uncertainty and inference for verification measures. *Weather Forecast.* **22**: 637–650.
- Keil C, Craig GC. 2007. A displacement-based error measure applied in a regional ensemble forecasting system. *Mon. Weather Rev.* **135**(9): 3248–3259.
- Keil C, Craig GC. 2009. A displacement and amplitude score employing an optical flow technique. *Weather Forecast.* **24**(5): 1297–1308.
- Keogh E, Ratanamahatana CA. 2004. Exact indexing of dynamic time warping. *Knowl. Inf. Syst.* **7**: 358–386.
- Lettau M, Ludvigson S. 2001. Consumption, aggregate wealth and expected stock returns. *J. Finance* **56**(3): 815–849.
- Lin J, Cervone G, Franzese P. 2010. Assessment of error in air quality models using dynamic time warping. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Data Mining for Geoinformatics (DMG '10)* 2 November 2010. ACM: New York, NY; 38–44. DOI: 10.1145/1869890.1869895.
- Liu Y, Hopson T, Roux G, Hacker J, Wu W, Warner T, et al. 2009. An operational mesoscale ensemble data assimilation and prediction system: E-RTFDAA – system design and verification. In *19th Conference of Numerical Weather Prediction/23rd Conference of Weather Forecasting*, 1–5 June 2009, Omaha, NE.
- McCracken MW. 2004. Parameter estimation and tests of equal forecast accuracy between non-nested models. *Int. J. Forecast.* **20**: 503–514.
- Marzbán C, Sandgathe S. 2010. Optical flow for verification. *Weather Forecast.* **25**(5): 1479–1494.
- Myers C, Rabiner L, Rosenberg AE. 1980. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Trans. Acoust. Speech Signal Process.* **28**(6): 623–635.
- NCAR - Research Applications Laboratory. 2012. Verification: forecast verification utilities. R package version 1.37. <http://CRAN.R-project.org/package=verification>. (accessed 1 May 2014)
- Nehrkorn T, Hoffman RN, Grassotti C, Louis J-F. 2003. Feature calibration and alignment to represent model forecast errors: empirical regularization. *Q. J. R. Meteorol. Soc.* **129**(587): 195–218.
- Poon S-H, Granger CWJ. 2003. Forecasting volatility in financial markets: a review. *J. Econ. Lit.* **41**(2): 478–539.
- R Core Team. 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna ISBN 3-900051-07-0 <http://www.R-project.org/> (accessed 1 May 2014).
- Rabiner L, Rosenberg A, Levinson S. 1978. Considerations in dynamic time warping algorithms for discrete word recognition. *IEEE Trans. Acoust. Speech Signal Process.* **26**: 575–582.
- Sakoe H, Chiba S. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Process.* **26**: 43–49.
- West KD. 1996. Asymptotic inference about predictive ability. *Econometrica* **64**: 1067–1084.
- Wilks DS. 1997. Resampling hypothesis tests for autocorrelated fields. *J. Clim.* **10**: 65–82.