From the Desk

March 2010

Confidence Intervals for Forecast Verification: Practical Considerations

Eric Gilleland

Research Applications Laboratory

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH
P. O. Box 3000
BOULDER, COLORADO 80307-3000

Contents

PREFACE **v**

1 Introduction **1**

2 Parametric vs. Non-parametric Confidence Intervals **3**
2.1 Which should I use? 4

3 Dependence assumption **6**
3.1 Checking for dependence 7

4 Between the wheels: Deciding how to apply confidence intervals for multiple fields of a forecast model **10**

5 How confident do we want to be? **17**

6 What is simultaneous confidence? **17**

7 How do assumption violations affect confidence intervals? An *ad-hoc* study using simulations **17**

References **29**

List of Figures

1 *Forecast bias (with 99% confidence intervals) for 24-hour accumulated precipitation (in) calculated for increasing thresholds for ARW (blue), NMM (red) and the difference of NMM subtracted from ARW (green). Figure reproduced from Bernadet et al. (2009).* 3

2 Sample autocorrelation- (top) and partial autocorrelation- (bottom) function graphs for a simulated white noise process. Dashed blue lines represent $\pm 1.96/\sqrt{1000}$. Data were simulated in R (R Development Core Team, 2008) using `arima.sim( list(), n=1000)`, and the `acf` and `pacf` functions were used to create the graphs shown here, as well as all subsequent autocorrelation- and partial autocorrelation graphs in this document. 11

3 Sample autocorrelation- (top) and partial autocorrelation- (bottom) function graphs for a simulated AR(2) process. Dashed blue lines represent $\pm 1.96/\sqrt{1000}$. Data were simulated in R (R Development Core Team, 2008) using `arima.sim( list( order=c(2,0,0), ar=c(0.7,-0.5)), 1000)`. 12

4 Sample autocorrelation- (top) and partial autocorrelation- (bottom) function graphs for a simulated MA(1) process. Dashed blue lines represent $\pm 1.96/\sqrt{1000}$, and dashed red lines show the bounds $\pm 1.96/\sqrt{1000}\sqrt{1+2\hat{\rho}^2(1)}$. Data were simulated in R (R Development Core Team, 2008) using `arima.sim(list( order=c(0,0,1), ma=0.8), 1000)`. 13

5 Sample autocorrelation- (top) and partial autocorrelation- (bottom) function graphs for a simulated ARIMA(1,1,1) process. Dashed blue lines represent $\pm 1.96/\sqrt{1000}$. Data were simulated in R (R Development Core Team, 2008) using `arima.sim(list( order=c(1,1,1), ar=0.75, ma=0.6), 1000)`. 14

6 Normal qq-plot for 1000 simulated mean errors. 20

7 Normal qq-plots for 1000 simulated frequency bias statistics, using thresholds of two, five, eight, eleven and fourteen (from upper left corner moving from left to right in each row successively). 21

8 Normal qq-plot for 1000 simulated root mean squared errors. 22

9 Normal qq-plots for 1000 simulated hit rate statistics, using thresholds of two, five, eight, eleven and fourteen (from upper left corner moving from left to right in each row successively). 23

10 Simulation of size 1000 from an AR(2) model with $\rho_1 = 0.7$ and $\rho_2 = -0.4$. Top left is time series, top right is the normal qq-plot, and autocorrelation- and partial autocorrelation function graphs are shown in the bottom row. 25

List of Tables

1 Assumptions implicit in standard confidence intervals for a statistic θ estimated by $\hat{\theta}$ from a sample $x_1, \ldots, x_n$ by method. 5

2 95% confidence intervals for some simulated verification statistics from one sample of the 1000 simulated samples. Standard errors for rmse and bias are estimated by the standard deviations of these statistics across all 1000 samples. Hit rate normal approximation confidence intervals are calculated using Wald's interval (see Gilleland, 2010, section 2.8.1 Eq (11)). 19

3 IID Bootstrap (BC_a) 95% confidence intervals for verification statistics from two simulated forecasts against the same verification series, and for the differences between statistics for each forecast. . . 26

4 95% confidence intervals for the mean (≈ 0.03) of the simulated series shown in Figure 10. Normal approximation with variance inflation factor assumes aan AR(1) model. 27

PREFACE

This manuscript is intended to give a simple overview to constructing and using confidence intervals for the purposes of forecast verification. A more detailed manuscript on the subject can be found in Gilleland (2010).

1 Introduction

Traditional forecast verification summarizes forecast performance through statistics such as the root mean square error (continuous fields), probability of detection (contingency tables), etc. When obtaining such a point estimate of a random quantity the value observed is a single realization from a distribution of possible values. Therefore, in order to make inferences about the true value of the quantity, say θ, it is important to acknowledge this uncertainty about the point estimate. A common and useful way of doing this is to report a confidence interval for the quantity $\hat{\theta}$. Under the classical (or frequentist) paradigm, a $(1-\alpha)\cdot 100\%$ confidence interval for a parameter estimate $\hat{\theta}$ is interpreted, awkwardly, as the interval such that if it were reconstructed for 100 different realizations of the sample of random variables, one would expect that the true parameter, θ, would fall inside $(1-\alpha)\cdot 100$ of the intervals.

Confidence intervals are treated in any introductory statistical textbook, and specifically for literature applied to forecast verification, see for example Jolliffe (2007), Wilks (2006), Jolliffe and Stephenson (2003) and Gilleland (2010). This write-up is an attempt to consolidate as much pertinent information on confidence intervals as is reasonable (particularly as pertains to forecast verification) into a single manuscript. For Bayesian Credible Intervals, see for example, Bernardo and Smith (2000) and Jolliffe (2007). Fiducial confidence intervals as originally argued for by Fisher (1935) is another frequentist approach intended to provide frequentist intervals with a more favorable philosophical foundation to compete with Bayesian intervals (e.g., Dawid and Stone, 1982; Hannig et al., 2006), confidence intervals for this and the Bayesian paradigm are not discussed here. Another type of frequentist confidence interval not discussed here is the profile-likelihood method (e.g., Coles, 2001; Gilleland and Katz, 2006) as this approach is not typically useful for most forecast verification purposes.

Throughout this write-up, f will denote the forecast series (or field) and o the verification series (or field), which may be a direct observation, or an interpolation or analysis. Indeed, the validity of confidence intervals may be affected by the nature of the verification series. For example, if multiple forecasts are being compared, and an analysis obtained from one of the forecasts, say forecast A, is used as the verification series, then there may be an artificial bias favoring forecast A over the others. Further, even if direct observations are taken, there is inevitably an error associated with them. Observational error can be an important component of uncertainty that should be incorporated into the confidence intervals. However, this issue is not discussed further here.

The following example will be used throughout as it illustrates a common analysis required for many evaluations of forecast performance. It also illustrates many of the points highlighted in this manuscript.

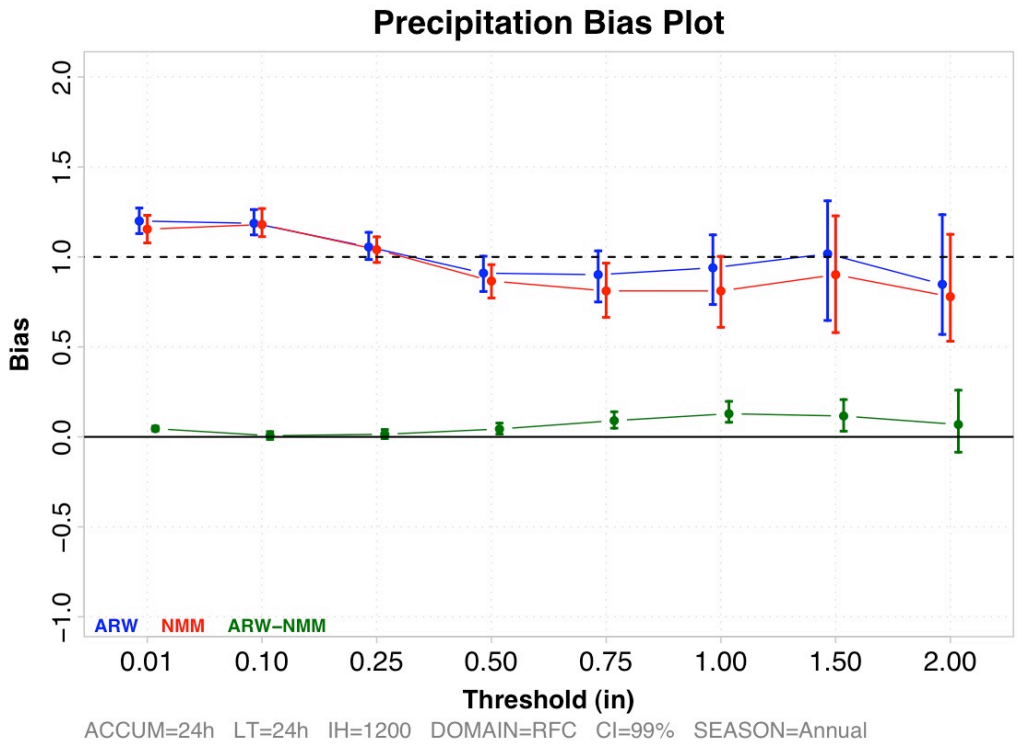
Example 1.1. Figure 1 is an example of confidence intervals obtained from bootstrapping for two versions of the WRF model for 24-hour accumulated precipitation (in) for increasing thresholds. The bias values were taken from adequately spaced points in time so that temporal dependence is not considered to be an issue.

Meaningful arguments about forecast performance can be made from these intervals. For example, the uncertainty increases as the threshold increases. In this case, the source of the increased uncertainty is most likely from having fewer observations exceeding the higher thresholds. Generally, it will be more difficult to make strong conclusions in the face of such large uncertainty. For the lower thresholds, where the bias estimates are more certain, it can be seen that for both models, the value of one does not lie within the intervals; the estimate and bounds are all greater than one. The interpretation for these thresholds, then, is that with 99% confidence both models are over-forecasting precipitation above 0.01 and 0.10 inches. For other thresholds, it cannot be strongly argued that the ARW model is not unbiased.

While tempting to draw the conclusion from the red and blue intervals in the figure that the ARW and NMM models are not significantly different from each other in terms of bias because their intervals overlap, this is not a defensible conclusion. To make such inferences, we need to observe the bias differences for between the two models as shown by the green line in the figure. Upon doing so, we see that there are thresholds where the ARW model is statistically significantly, with 99% confidence, more biased than the NMM model, despite that their confidence intervals overlap.

Note that this is not a statement about both models simultaneously, or about both thresholds together (see section 2.12 in Gilleland, 2010). Rather, these intervals were constructed individually for each stratification of threshold, model, season, domain, initialization- and lead- times, etc. Therefore, we cannot draw conclusions about model perform-

Figure 1: *Forecast bias (with 99% confidence intervals) for 24-hour accumulated precipitation (in) calculated for increasing thresholds for ARW (blue), NMM (red) and the difference of NMM subtracted from ARW (green). Figure reproduced from Bernadet et al. (2009).*



ance across thresholds; though this is a fine point for the present example.

2 Parametric vs. Non-parametric Confidence Intervals

There are many ways to construct valid confidence intervals, and generally these can be divided into parametric and non-parametric intervals. By far the most common parametric intervals are the normal approximation intervals, and several others require an assumption of normality for the underlying sample (e.g., con-

fidence intervals for variance and correlation (cf. Gilleland, 2010, sections 2.5 and 2.6)). The $(1-\alpha)\cdot 100\%$ normal approximation confidence interval for an estimate $\hat{\theta}$ of the true value of some statistic θ is given by

$$\hat{\theta} \pm z_{1-\alpha/2}\mathrm{se}(\theta), \tag{1}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution function (df), which because the normal df is symmetric, we have that $z_{1-\alpha/2} = z_{\alpha/2}$. Note that to obtain a 95% normal approximation confidence interval, the 2.5 and 97.5 quantiles of the standard normal df are needed. This is because 95% of the mass of the normal df falls between these two quantiles (i.e., $1-\alpha/2-\alpha/2 = 1-\alpha$, so for $\alpha = 0.05$, $1-\alpha = 0.95$).

The most common type of non-parametric confidence interval is the bootstrap interval. This manuscript will focus only on normal approximation and bootstrap intervals. Briefly, a bootstrap interval is obtained from a sample of the statistic of interest, $\theta_1, \ldots, \theta_B$. Such a sample is obtained from re-estimating θ numerous times, say B times, from random samples of the data with replacement. There are several ways to estimate the intervals from the sample $\theta_1, \ldots, \theta_B$ (see Gilleland, 2010, section 3.2), but here we will only consider the percentile and BC_a intervals. The percentile interval is straightforward and easy to compute, but has stronger assumptions than the BC_a interval that, if violated, will yield too narrow intervals. The BC_a will be identical to the percentile intervals if the assumptions are met, and are otherwise more accurate. The drawback to the BC_a interval is that it requires another round of re-sampling making it computationally more burdensome. While the assumptions for bootstrap confidence intervals are considerably less rigid than those for the normal intervals, there is still an assumption that the data are independent and identically distributed (iid) and that the sample is representative of the population. Therefore, the bootstrap procedure just described will be referred to as the iid bootstrap in order to emphasize this important point, and to distinguish it from other bootstrap procedures that do not require this assumption. Table 1 summarizes the assumptions implied by each method (see Gilleland, 2010, section 3 for more about the other bootstrap types).

2.1 Which should I use?

Many statistics, including many verification statistics, can take the normal approximation interval, or one derived from a normality assumption. The assumption of normality should be checked (see Gilleland, 2010, section 2.10), but when they can be computed, they are quick and easy to construct; provided a reasonable estimate of the standard error for the statistic can be found. See Gilleland (2010) section 2 for information about which verification statistics can take the normal approximation interval.

Table 1: Assumptions implicit in standard confidence intervals for a statistic θ estimated by $\hat{\theta}$ from a sample $x_1, \ldots, x_n$ by method.

Method	Assumptions
Normal Approximation	$\hat{\theta} \sim N(\theta, \mathrm{se}(\theta))$. $x_1, \ldots, x_n$ are iid. Symmetric interval (implied by first assumption). $\mathrm{se}(\theta)$ is known or can be well estimated by $\widehat{\mathrm{se}}(\theta)$.
IID Bootstrap (percentile)	$x_1, \ldots, x_n$ are iid. $\hat{\theta}$ follows a symmetric df with constant variance. The sample df of $x_1, \ldots, x_n$ is representative of the population df.
IID Bootstrap (BC_a)	$x_1, \ldots, x_n$ are iid. The sample df of $x_1, \ldots, x_n$ is representative of the population df.
Block Bootstrap	$x_1, \ldots, x_n$ are identically distributed. Block lengths are substantially longer than the dependence length. The sample df of $x_1, \ldots, x_n$ is representative of the population df.
Parametric Bootstrap	Parametric model is correct. Whatever assumptions there are about the parametric model are correct.

Even when the normal approximation interval is appropriate, bootstrap intervals may be better in terms of accuracy; they will not be worse. There are less rigid assumptions for them, and they do not require calculation (for most methods, such as the percentile and BC_a) of a standard error for the statistic. The drawback of the bootstrap procedure is that it can be computationally expensive. In terms of which type of bootstrap confidence intervals to use, the percentile is a good choice, but if the assumption of symmetry and constant variance for the statistic's df are violated, the intervals will be too narrow. The BC_a obviates these assumptions, and when they are met resuces to the percentile intervals. The drawback of this method is that it requires an additional resampling procedure to compute; making it less computationally efficient than the percentile method. Another method, called the ABC (Gilleland, 2010, section 3.2.5), is an approximation to the BC_a, but can only be computed for smooth statistics (e.g., the mean) but not for other types (e.g., the median).

3 Dependence assumption

In most cases it is reasonable to assume that the data follow the same df, or that any changes in distributional form can be easily handled in some way. For example, in figure 1.1 the confidence intervals are calculated for reasonably homogenous stratifications of verification fields (i.e., same season, same initialization/valid times, etc.). To check the assumption of normality (for normal approximation intervals) see Gilleland (2010) section 2.10.

The assumption about independence, however, is most often violated. The intervals shown in figure 1.1 are estimated from samples that were taken over relatively long intervals in time so that the independence assumption is reasonable. However, if the assumption were found to be invalid, then it would be expected that the intervals are too narrow because effectively there are *repeated* observations. That is, we have fewer sample points than it appears we have. Fewer sample points implies that there is greater uncertainty than has been accounted for. This would not affect conclusions pertaining to unbiasedness for those thresholds that encompass one, but it might affect conclusions drawn, for example, about the two lowest thresholds pertaining to unbiasedness. See section 2.11 in Gilleland (2010) for a simple method of accounting for dependence with normal approximation intervals; namely using the variance inflation factor. To account for dependence with the bootstrap procedure, the simplest method is to use block bootstrapping (Gilleland, 2010, section 3.4). If a reasonable model (e.g., an autoregressive model) exists to model the dependence, then a parametric bootstrap can be used, and such a scheme is generally preferable, especially in the face of spatial dependence. As mentioned in Table 1, the block bootstrap approach requires the block lengths

to be substantially longer than the dependence length. For time series, this is usually a reasonable assumption, but spatial dependence is often long relative to the spatial region making it a less reasonable assumption.

3.1 Checking for dependence

Statistically, two events A and B are independent if $\Pr\{A \text{ and } B\} = \Pr\{A\}\Pr\{B\}$, or equivalently, $\Pr\{A|B\} = \Pr\{A\}$ or $\Pr\{B|A\} = \Pr\{B\}$, where $\Pr\{A|B\}$ is read, "the probability of A given (or conditional on) B. If there is dependence in the two events, then $\Pr\{A \text{ and } B\} = \Pr\{A|B\}\Pr\{B\}$. If a sequence of random variables $X_1, \ldots, X_n$ is iid noise with df F and zero mean, then

$$\Pr\{X_1 \leq x_1, \ldots, X_n \leq x_n\} = \prod_{i=1}^{n} \Pr\{X_i \leq x_i\} = \prod_{i=1}^{n} F(x_i) = F^n(x_1).$$

In particular, if $g(X_1, \ldots, X_n)$ is a function that minimizes the mean squared error $E[(X_{n+h} - g(X_1, \ldots, X_n))^2]$, then g is identically zero.

Consider, for example, $E[(X_{i-1} - X_i)^2]$. Multiplying this out, we get

$$E[X_{i-1}^2] + E[X_i^2] - 2E[X_{i-1}X_i] = \mathrm{Var}X_{i-1} + \mathrm{Var}X_i - 2\mathrm{Cov}(X_{i-1}, X_i).$$

For an iid noise sequence, the last term is zero. Let $\gamma_X(h) = \mathrm{Cov}(X_{t+h}, X_t)$, then $\gamma_X(h)$ is called the autocovariance function of $\{X_t\}$, where $\gamma_X(0) = \mathrm{Var}X$. The autocorrelation function is given by

$$\rho_X(h) \equiv \frac{\gamma_X(h)}{\gamma_X(0)} = \mathrm{Cor}(X_{t+h}, X_t).$$

An iid noise sequence $\{X_t\}$, each with mean zero and variance σ^2 is referred to as white noise. A first-order moving average process is defined by

$$X_t = Z_t + \theta Z_{t-1},\ t = 0 \pm 1, \ldots \tag{2}$$

where the sequence $\{Z_t\}$ is white noise and θ is a real-valued constant. Such a process is dependent in time, whereby subsequent values are modified by a constant involving the previous time step. Note that

$$\mathrm{Cov}(X_{t+h}, X_t) = \begin{cases} \sigma^2(1+\theta^2) & h = 0 \\ \sigma^2\theta & h = \pm 1 \\ 0 & |h| > 1 \end{cases}$$

Higher order, say q-order, moving average processes are defined similarly, and have non-zero covariances between variables within the lag q of each other.

Another form of dependence pertains to the autoregression process. A first-order autoregression process is given by

$$X_t = \phi X_{t-1} + Z_t,\ t = 0, \pm 1, \ldots \tag{3}$$

where Z_t is again white noise with variance σ^2, $|\phi| < 1$. Higher-order autoregression processes are defined similarly. Note that now

$$\mathrm{Cov}(X_{t+h}, X_t) = \phi^h \gamma_X(0).$$

It is also possible to have more complicated dependence in the sequence (e.g., an autoregression moving average process), but the above will suffice for this treatment (in the spirit of keeping it as simple as possible). The reader is referred to Brockwell and Davis (2002) for more complex time series models.

When faced with data (i.e., without knowing what type of dependence model might be appropriate), a useful tool for determining what sort of dependence model might be appropriate is the sample autocorrelation function, $\hat{\rho}(h)$. The sample autocovariance function $\hat{\gamma}(h)$ for a sample $x_1, \ldots, x_n$ is defined by

$$\hat{\gamma}(h) := \frac{1}{n} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \bar{x})(x_t - \bar{x}),\ -n < h < n,$$

and the sample autocorrelation function $\hat{\rho}(h)$ is simply

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma(0)}},\ -n < h < n.$$

A graph depicting the autocorrelation function against increasing lags (e.g., $h = 0$ to $h = 40$) can be useful in diagnosing dependence features. For example, a sample of size n from an iid noise process follows a normal df with mean zero and variance $1/n$, for large n. Therefore, if a sample is from an iid noise process, then approximately $(1-\alpha)100\%$ of the sample autocorrelations should fall within the bounds $\pm z_{1-\alpha/2}/\sqrt{n}$. Therefore, in an autocorrelation graph with 40 lags, we expect $40(0.05) = 2$ values to fall outside these bounds. For a sample containing a trend $|\hat{\rho}|$ will display slow decay as h increases. If there is a substantial deterministic period, $|\hat{\rho}|$ will exhibit similar behavior with the same periodicity. for numerous other tests utilizing the sample autocorrelation function, refer to Brockwell and Davis (2002) section 1.6.

Although more complicated to define (see Brockwell and Davis, 2002, for the definition), the partial autocorrelation function is another useful tool for discerning autoregression and moving average models. Both the autocorrelation and partial autocorrelation functions can be easily graphed with any statistical software package. If a sample sequence is derived from a p-th order autoregression model, it

is expected that the first p lags h will be significantly different from zero, and negligible beyond $h = p$; this can be determined if about $n\alpha$ values beyond lag p to be outside the bounds $\pm z_{1-\alpha/2}/\sqrt{n}$.

Example 3.1. Graphical displays for assessing dependence.
For this example, simulations from samples of size 1000 each are constructed from different time series models, and autocorrelation and partial autocorrelation graphs diagnosed. Here, the R (R Development Core Team, 2008) function `arima.sim` is used to make all of the simulations.

Figure 2 shows these graphs for a white noise process. The autocorrelation function graph has a spike at lag zero indicating that the correlation of a value with itself is (correctly) unity. Fewer than one value beyond the zero lag exceeds the blue dashed lines. Further, no major departures from zero are present in the partial autocorrelation function graph, and fewer than $0.05(30) = 1.5$ points fall outside the dashed horizontal lines representing $\pm 1.96/\sqrt{1000}$ indicating that white noise is reasonable based on the autocorrelation function graph.

Figure 3 shows these graphs for a simulated AR(2) model. Greater than 1.5 values fall outside the horizontal blue lines in the autocorrelation function graph, and there are two large spikes significantly different from zero at the first two lags of the partial autocorrelation graph with fewer than 1.5 values outside the blue dashed lines beyond the first two lags. This indicates that the process is not iid noise, and suggests that an AR(2) model is reasonable for modeling the dependence in the time series.

Figure 4 shows the sample autocorrelation- and partial autocorrelation function graphs for a series that was simulated from an MA(1) process. The red dashed lines show the bounds $\pm 1.96/\sqrt{1000}\sqrt{1 + 2\hat{\rho}^2(1)}$, where $\hat{\rho}(1)$ is the sample autocorrelation at lag 1. Because $\hat{\rho}(h)$ is within these bounds for $h > 1$, it is reasonable to model the sample as an MA(1)

process. Note that the partial autocorrelation function graph suggests an AR(5) (or possibly higher order) model.

As a final example, Figure 5 shows these graphs for a simulated ARIMA(1,1,1) process. The slow decay in the autocorrelation graph suggests that an ARIMA process may be reasonable. In order to diagnose a specific appropriate model, differencing operations would need to be applied, and autocorrelation- and partial autocorrelation function graphs created for their residuals. Nevertheless, in each of Figures 3 through 5, the samples are clearly not independent. For the more straightforward cases (e.g., that of Figure 3), one could reasonably employ a parametric bootstrap procedure to obtain confidence intervals that account for dependence. Alternatively, in any of these cases, the lag dependence is clearly smaller than $\lfloor\sqrt{1000}\rfloor = 31$, and 31 is substantially smaller than the size of the sample ($n = 1000$) so that a block bootstrap could also be employed in this case. For the cases where an autoregression is a reasonable model, one can employ the variance inflation factor (see Gilleland, 2010, section 2.11) to obtain normal approximation confidence intervals that reasonably reflect the dependence in the sample.

4 Between the wheels: Deciding how to apply confidence intervals for multiple fields of a forecast model

Often in a forecast verification setting, there are multiple verification fields (e.g., over time) so that statistics (e.g., root mean squared error (rmse), contingency tables and their summary statistics, etc.) are often calculated for each field. That is, for example, a time series of contingency tables are generally available.[1] A question that arises pertains to how the confidence intervals should be calculated exactly. For example, they could be calculated for each time point (e.g., one could imagine a time series of hit rate with vertical bars at each point showing the

[1] This could be thought of as a time series of vectors containing the pertinent information from the contingency tables from which any contingency table derived statistic can be calculated.

Figure 2: Sample autocorrelation- (top) and partial autocorrelation- (bottom) function graphs for a simulated white noise process. Dashed blue lines represent $\pm 1.96/\sqrt{1000}$. Data were simulated in R (R Development Core Team, 2008) using `arima.sim( list(), n=1000)`, and the `acf` and `pacf` functions were used to create the graphs shown here, as well as all subsequent autocorrelation- and partial autocorrelation graphs in this document.

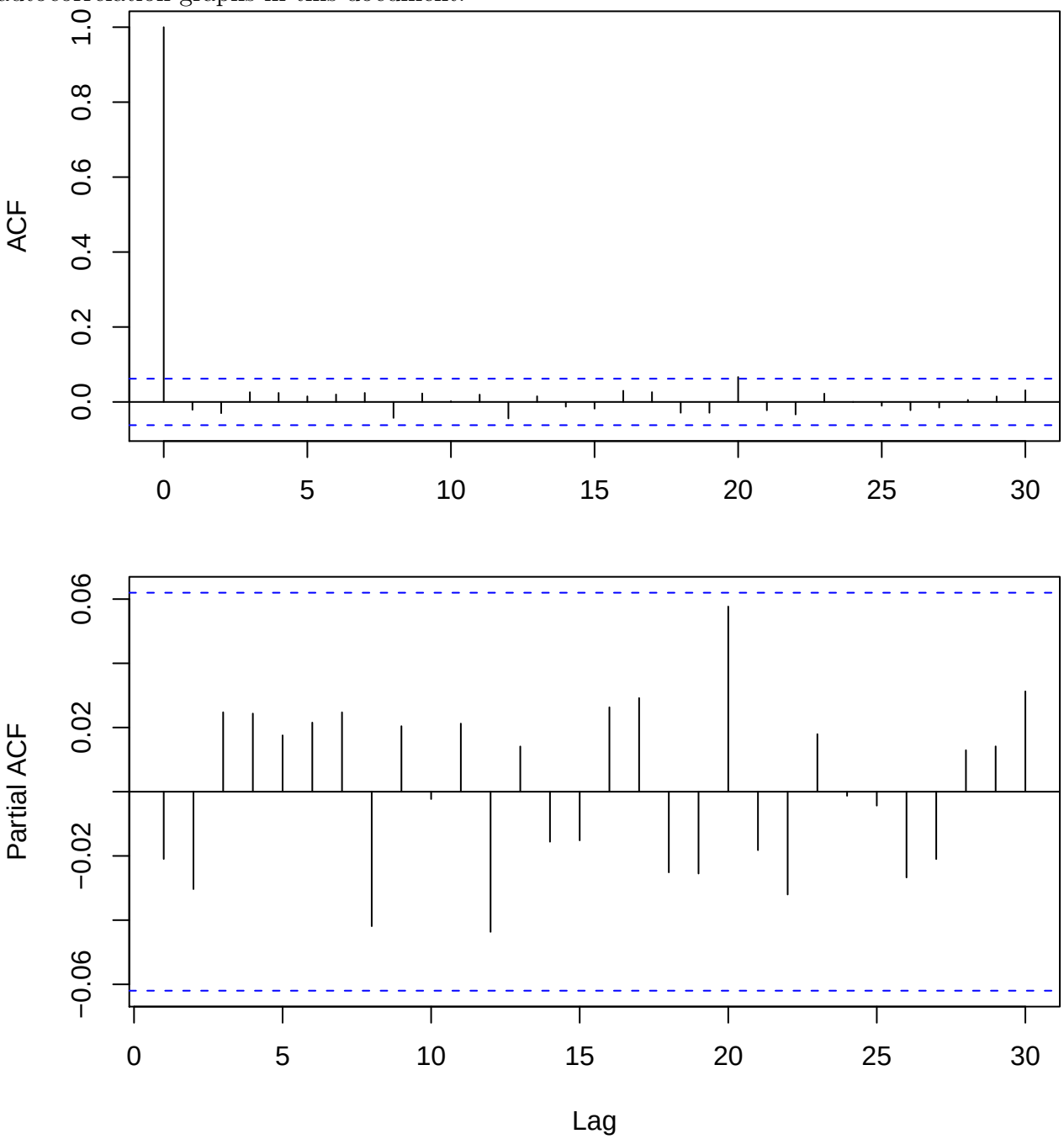


Figure 3: Sample autocorrelation- (top) and partial autocorrelation- (bottom) function graphs for a simulated AR(2) process. Dashed blue lines represent $\pm 1.96/\sqrt{1000}$. Data were simulated in R (R Development Core Team, 2008) using `arima.sim( list( order=c(2,0,0), ar=c(0.7,-0.5)), 1000)`.

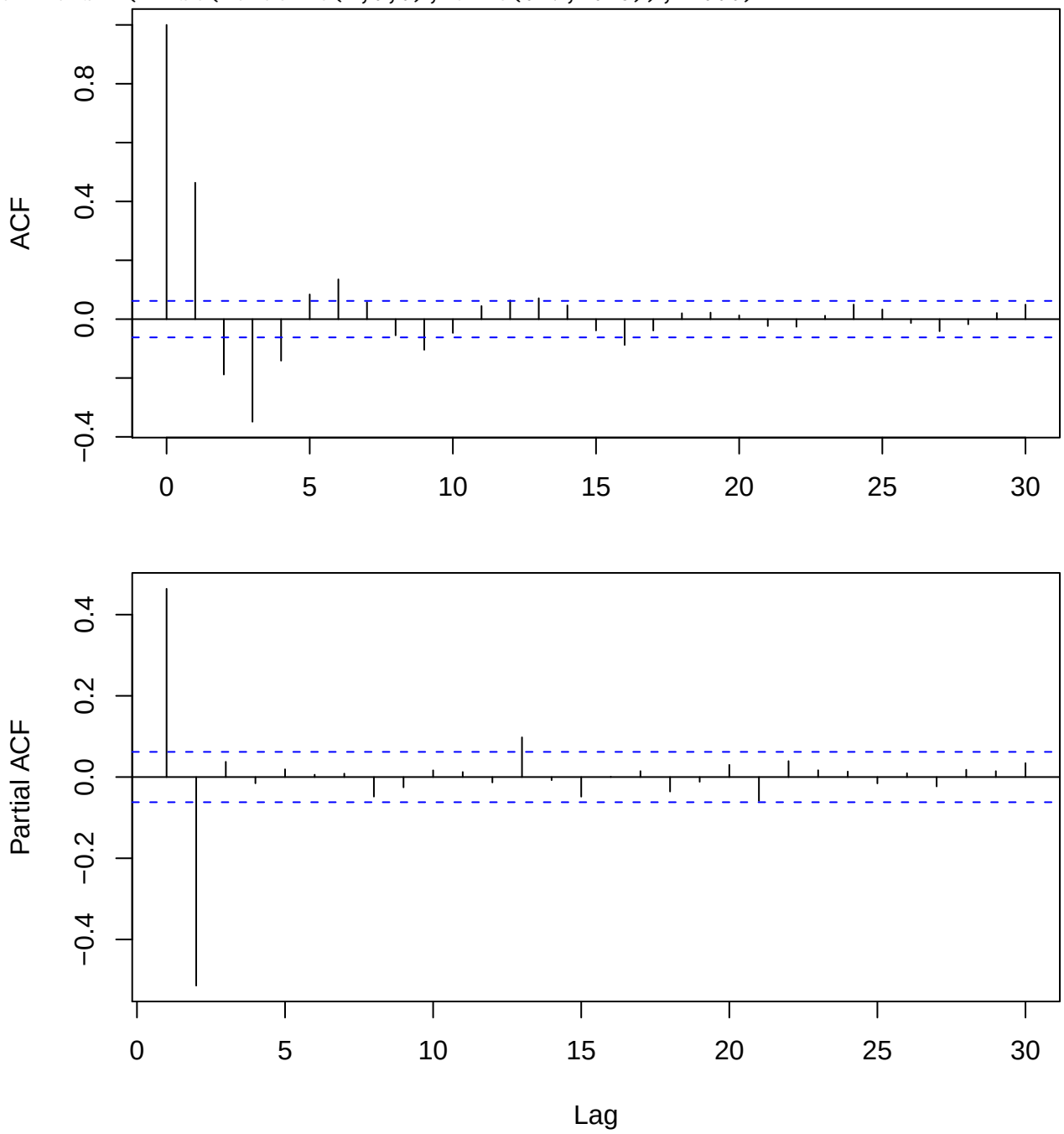


Figure 4: Sample autocorrelation- (top) and partial autocorrelation- (bottom) function graphs for a simulated MA(1) process. Dashed blue lines represent $\pm 1.96/\sqrt{1000}$, and dashed red lines show the bounds $\pm 1.96/\sqrt{1000}\sqrt{1+2\hat{\rho}^2(1)}$. Data were simulated in R (R Development Core Team, 2008) using `arima.sim(list( order=c(0,0,1), ma=0.8), 1000)`.

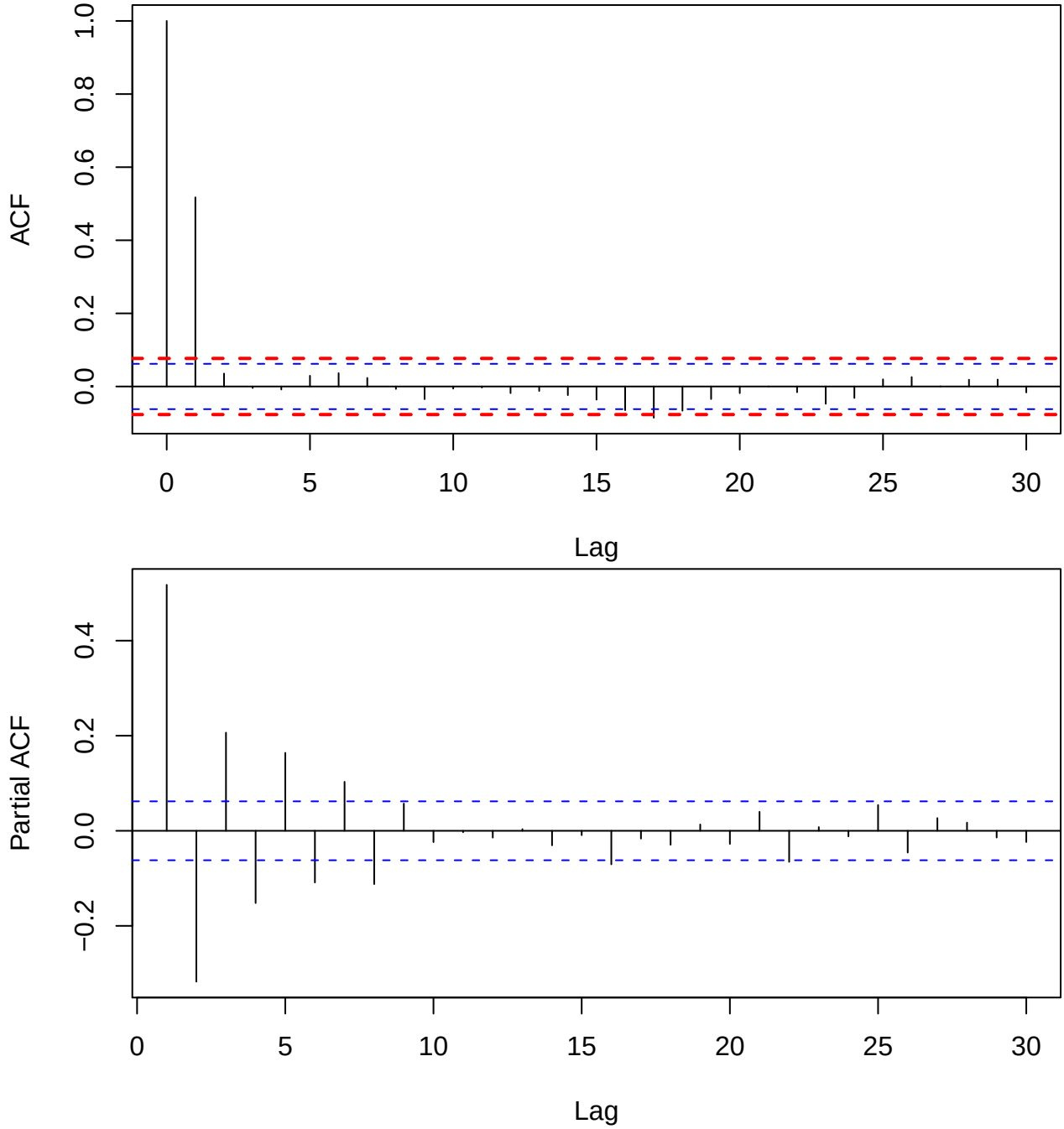
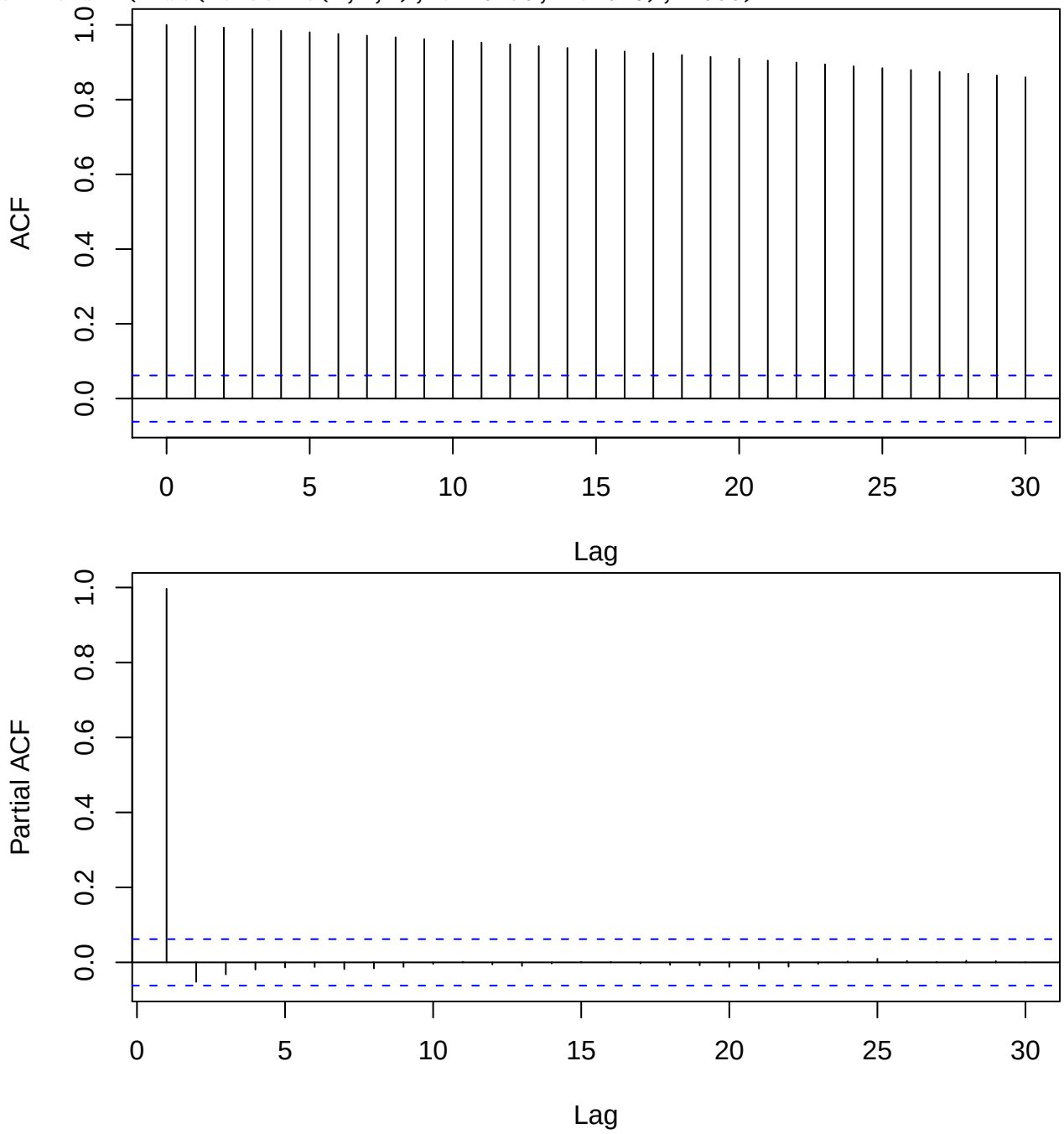


Figure 5: Sample autocorrelation- (top) and partial autocorrelation- (bottom) function graphs for a simulated ARIMA(1,1,1) process. Dashed blue lines represent $\pm 1.96/\sqrt{1000}$. Data were simulated in R (R Development Core Team, 2008) using `arima.sim(list( order=c(1,1,1), ar=0.75, ma=0.6), 1000)`.



confidence intervals), or they could be calculated on the average of the statistics across time, or they could be calculated on an aggregation of the tables across time.

Figure 1.1 shows confidence intervals for an aggregation of contingency tables over time. That is, at each iteration of resampling, the total number of forecast events are added across tables, and the total number of observed events are added across tables, and the bias is determined from these values. That is, for $j = 1, \ldots, n$ contingency tables, the estimated bias, call it $\hat{\theta}$, at the i-th iteration is determined by

$$\hat{\theta}_i = \frac{\sum_{j=1}^{n} \{\text{Number of forecast events}\}_j}{\sum_{j=1}^{n} \{\text{Number of observed events}\}_j}.$$

Note that this differs from calculating confidence intervals for the average bias across time. That is, in general

$$\frac{\sum_{j=1}^{n} \{\text{Number of forecast events}\}_j}{\sum_{j=1}^{n} \{\text{Number of observed events}\}_j} \neq \frac{1}{n} \sum_{j=1}^{n} \frac{\{\text{Number of forecast events}\}_j}{\{\text{Number of observed events}\}_j}.$$

Confidence intervals for the left hand side of the above equation inform about the uncertainty pertaining to bias across time, or the *between-field uncertainty*, whereas the right hand side informs about uncertainty pertaining to the average bias across time. Either is correct, but each informs about different types of uncertainty. When calculated for each field separately, the intervals inform about the *within-field uncertainty*.

Example 4.1. Beneath, between and behind
Within-field uncertainty is calculated for a single field (i.e., at a single point in time), which can be useful to determine if the uncertainty is changing over time (e.g., this might indicate some change in the nature of the field, such as fewer available values). Between-field uncertainty does not inform about how the uncertainty is evolving over time (i.e., the value of the statistic for a given field is assumed fixed), but rather informs about the uncertainty in the aggregated value of the statistic over time. Average uncertainty, while similar

to between-field uncertainty, informs about the uncertainty of the average value of the statistic of interest.

Suppose a verification set (i.e., a forecast f and observation/analysis o) is available for m time points. Then, we have $f_1, \ldots, f_m$ paired with $o_1, \ldots, o_m$. For simplicity, say interest is in the mean error $\overline{f-o}$. Suppose further that $\varepsilon_{1j} = f_{1j} - o_{1j}, \ldots, \varepsilon_{nj} = f_{nj} - o_{nj}$ are iid normally distributed for each verification set j. For a given field, the mean error is estimated by $\bar{\varepsilon}_j = \sum_{i=1}^{n} \varepsilon_{ij}/n$, and the variance (of errors, ε_j) by $\hat{\sigma}^2_{\varepsilon,j} = \sum_{i=1}^{n} (\varepsilon_{ij} - \bar{\varepsilon}_j)^2/(n-1)$.

Within-field $(1-\alpha) \cdot 100\%$ normal approximation confidence intervals can be constructed for each field ε_j, $j = 1, \ldots, m$ by

$$\bar{\varepsilon}_j \pm z_{1-\alpha/2} \hat{\sigma}_{\varepsilon,j}/\sqrt{n}.$$

Between-field $(1-\alpha) \cdot 100\%$ normal approximation confidence intervals can be constructed by

$$\bar{\varepsilon} \pm z_{1-\alpha/2} \hat{\sigma}_{\varepsilon}/\sqrt{nm},$$

where $\bar{\varepsilon} = \sum_{i=1}^{n} \sum_{j=1}^{m} \varepsilon_{ij}/(nm)$ and $\hat{\sigma}^2_{\bar{\varepsilon}} = \sum_{i=1}^{n} \sum_{j=1}^{m} (\varepsilon_{ij} - \bar{\varepsilon})^2/(nm-1)$.

Finally, $(1-\alpha) \cdot 100\%$ normal approximation confidence intervals for the average of the mean errors is constructed by

$$\bar{\varepsilon} \pm z_{1-\alpha/2} \hat{\sigma}_{\bar{\varepsilon}}/\sqrt{m},$$

where $\bar{\epsilon} = \sum_{j=1}^{m} \bar{\varepsilon}_j/m$ and $\hat{\sigma}_{\bar{\varepsilon}} = \sum_{j=1}^{m} (\bar{\varepsilon}_j - \bar{\varepsilon})^2/(m-1)$. Notice that for mean error, the estimated between-field statistic is the same as that for the average, but that their variance estimates differ.

5 How confident do we want to be?

A question often asked of statisticians is, “what level α should I use?" Statisticians are wont to reply with, “you tell me." In fact, there is no right or wrong answer to the question, but ultimately it depends on how precise one needs to be. The lower the confidence level, the narrower the interval because we are less confident that the true parameter will be inside the interval. Higher confidence levels will lead to wider intervals because we are more confident that the true parameter will be within the interval. When unsure about what level to use, it is customary to report intervals for a few levels (e.g., 95%, 99% and 99.9%) to get a feel for the sensitivity of the results.

6 What is simultaneous confidence?

When interest is in the joint uncertainty of more than one statistic, or across time points, spatial locations, etc., then individually calculated intervals will generally be too narrow to make inferences about the joint values. Most software packages that calculate notches on box plots, giving typically the 95% confidence intervals about the median, incorporate a confidence level that approximates a joint uncertainty so that the medians can be compared for pairs of box plots (see Gilleland, 2010, section 2.4 for more about the normal approximation confidence interval for the median).

The issue is related to multiple testing when conducting statistical hypothesis tests, for example, at many grid locations. If a 5% level test is passed at 1000 individual locations, it is expected that 50 of those tests are wrong. Therefore, it is not valid to speak about regions of significance without accounting for this issue.[2] See Gilleland (2010) section 2.12 for more about simultaneous confidence intervals including a simple method for calculating them.

7 How do assumption violations affect confidence intervals? An *ad-hoc* study using simulations

In this section, we will simply simulate many samples from each of several types of situations, and apply different statistics to those samples. Confidence intervals will be obtained using both the normal approximation (using the standard deviation of the various sample statistics to derive a standard error approximation for statistics without easily obtained estimates) and the bootstrap (BC_a) method.

[2] In this context, there is also likely an issue with spatial dependence.

Using the techniques described in Gilleland (2010) appendix A.3, we first simulate two forecasts and one observation triplets (i.e., three correlated random vectors) to obtain 1000 samples of size 1000 each, for each triplet. In each case, a baseline mean of ten and standard deviation of two are allowed to vary for each sample (and each member of the triplet individually) by the addition of a standard normal random number for the mean, and a random uniform value from 1×10^{-8} to three for the standard deviation. Additionally, the amount of correlation is allowed to vary according to a uniform df from one half to one, but the correlation between one forecast and the observed vector is the same as that between the other forecast and the observation for each triplet (this is done for simplicity). For this first group, no temporal correlation is added, so it is reasonable to expect that the samples are iid. the first question we want to examine is how well various verification statistics adhere to the assumption of normality.

It is expected that the mean error would follow a normal df. Figure 6 shows a normal qq-plot for the simulated mean errors across the 1000 samples for one of the two forecasts. The points lie in a reasonably straight line suggesting that the assumption of normality is reasonable for this statistic. The normal qq-plot for the simulated frequency bias statistics for these same samples, on the other hand, shows departures from the straight line, indicating that the normality assumption is not valid for the bias statistic (Figure 7). How far the departure is from normality for bias depends greatly on the threshold. For the simulations here, the lowest threshold of two shows an S-shaped curvature, but otherwise relatively straight; indicating that the df for bias using a threshold of two for these simulations is skewed or has heavier tails than the normal df. The two moderately low thresholds (five and eight) show a concave curvature, where the lower of the two is nearly straight. The highest two thresholds show a high amount of concavity, and are clearly not normally distributed.

Figure 8 shows the normal qq-plot for the sample of 1000 rmse values. Given that the rmse is an average value, a normality assumption is not out of the question. As the qq-plot is relatively straight, such an assumption may be valid for this statistic. The difficulty for calculating normal approximation confidence intervals for rmse, however, concerns estimation of its standard error.

Finally, we take a look at the hit rate for the same thresholds as used for frequency bias (Figure 9). Of concern here is that theoretically the hit rate should be approximately normal because $n = 1000$ is reasonably large and the statistic is a proportion (Gilleland, 2010, section 2.8.1). However, for all of the thresholds, the normal qq-plots exhibit severe departures from normality.

Because the confidence intervals for variance, standard deviation and correlation are all derived from the underlying sample's being iid normally distributed, but do not use the usual normal approximation interval, they are not investigated

Table 2: 95% confidence intervals for some simulated verification statistics from one sample of the 1000 simulated samples. Standard errors for rmse and bias are estimated by the standard deviations of these statistics across all 1000 samples. Hit rate normal approximation confidence intervals are calculated using Wald's interval (see Gilleland, 2010, section 2.8.1 Eq (11)).

Statistic (estimate)	**Normal approximation**	**Bootstrap (BC_a)**
Mean Error (≈ -1.51)	$(-1.69, -1.33)$	$(-1.69, -1.34)$
rmse (≈ 3.24)	(1.49, 4.99)	(3.12, 3.39)
Bias (threshold=5) (≈ 0.84)	(0.65, 1.03)	(0.82, 0.87)
Bias (threshold=14) (≈ 2.41)	(−8.99, 13.82)	(2.01, 2.92)
Hit rate (threshold=5) (≈ 0.84)	(0.82, 0.87)	(0.82, 0.87)
Hit rate (threshold=14) (≈ 0.93)	(0.92, 0.95)	(0.86, 0.98)

here.

Table 2 normal approximation and bootstrap (BC_a) confidence intervals for mean error, rmse, bias and hit rate for one of the 1000 simulated forecast-observation pairs. Of particular note is that the intervals for mean error (which should follow a normal df) are nearly identical between the two methods. On the other hand, the rmse intervals differ wildly, despite that the normal qq-plot for rmse based on the simulated samples follows closely to a straight line (Figure 8). In this case, the problem could be the estimate for the standard error of rmse, which was taken here to be the standard deviation of the rmse values obtained across the 1000 samples (≈ 0.89), and may be too large. For both bias statistics (using thresholds of five and fourteen), the intervals differ greatly, where the normal intervals are too wide

Figure 6: Normal qq-plot for 1000 simulated mean errors.

Figure 7: Normal qq-plots for 1000 simulated frequency bias statistics, using thresholds of two, five, eight, eleven and fourteen (from upper left corner moving from left to right in each row successively).

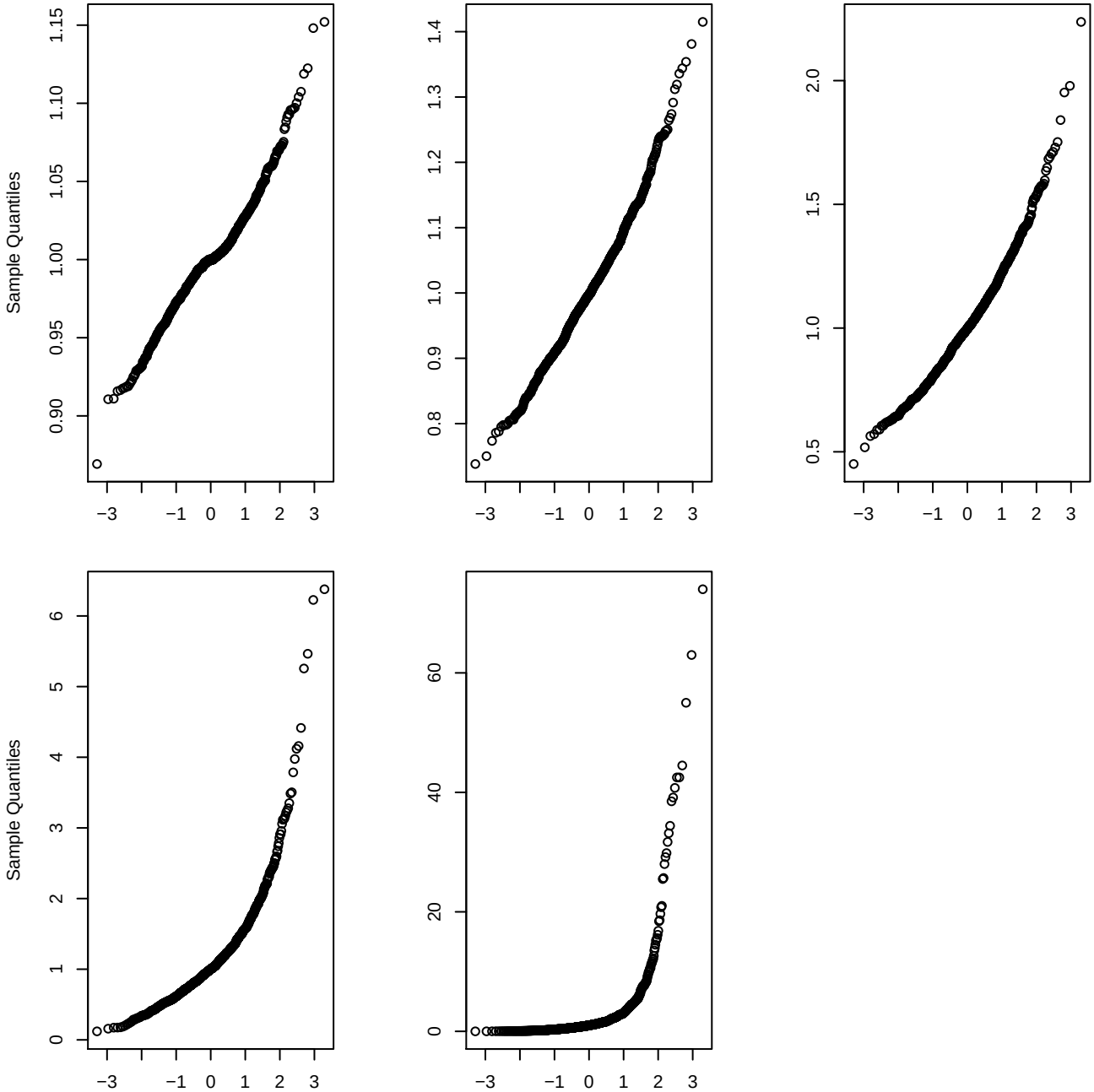


Figure 8: Normal qq-plot for 1000 simulated root mean squared errors.

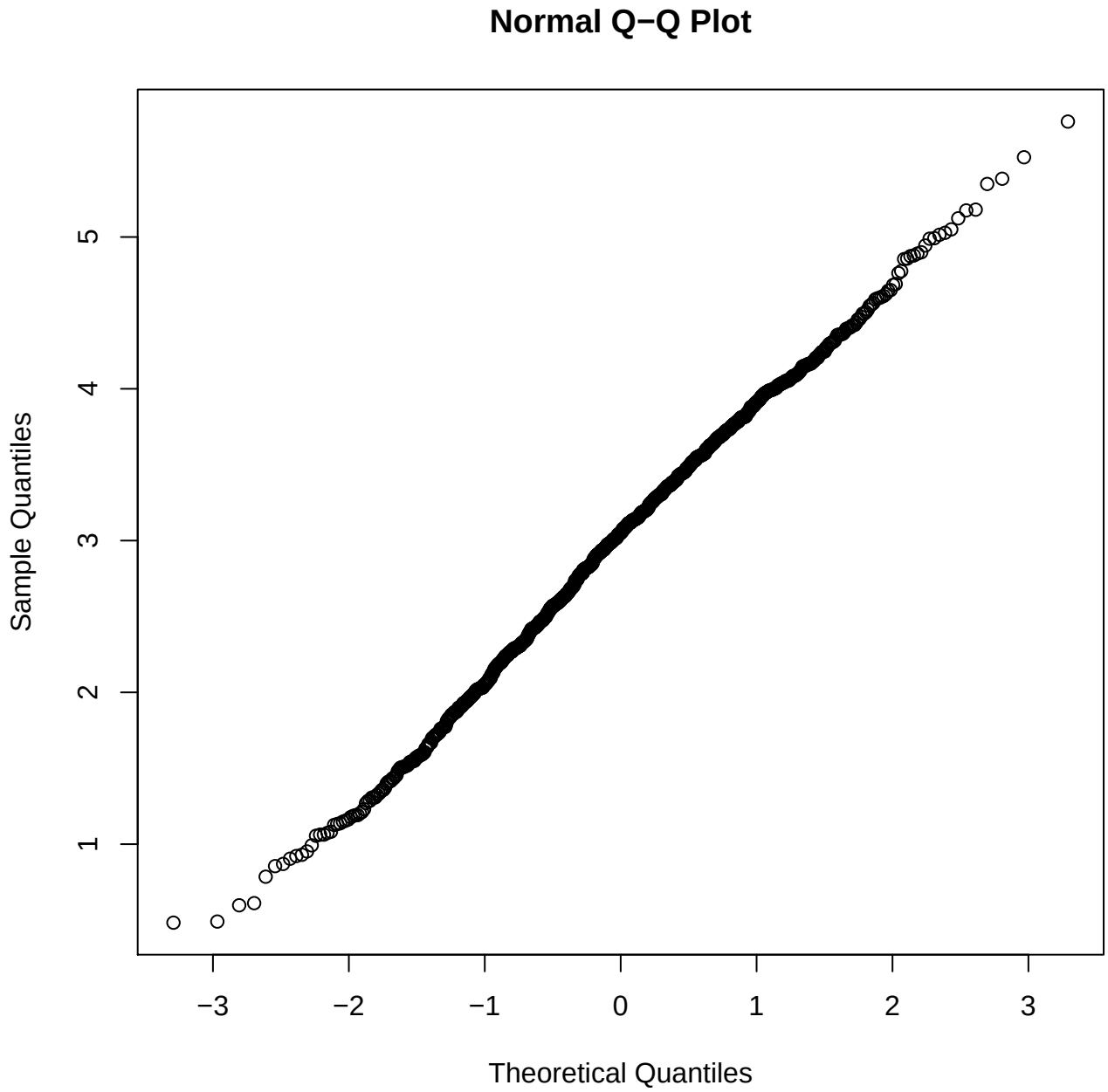
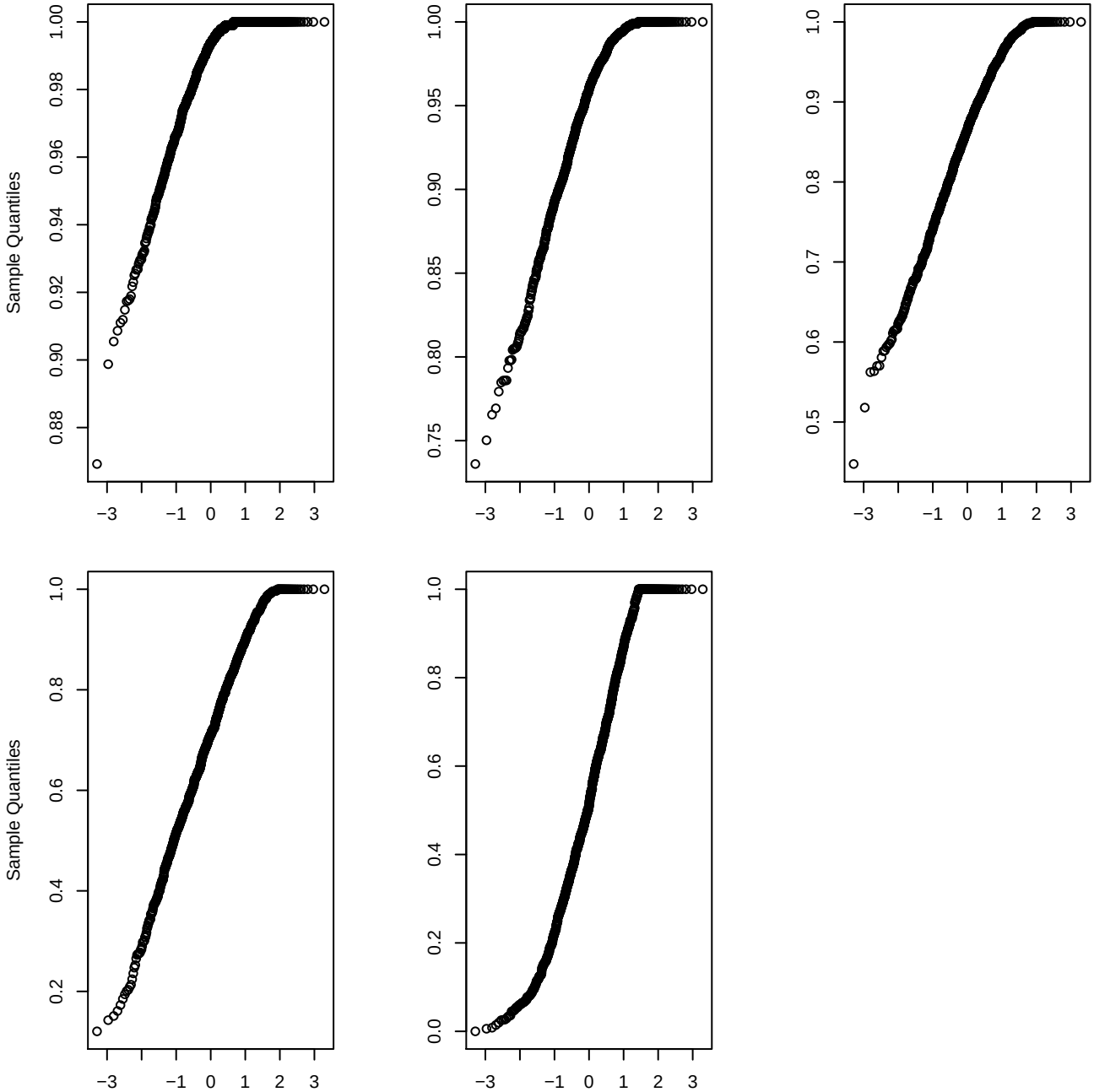


Figure 9: Normal qq-plots for 1000 simulated hit rate statistics, using thresholds of two, five, eight, eleven and fourteen (from upper left corner moving from left to right in each row successively).



and unrealistically symmetric. It is also interesting to note that for the threshold of five, the bias estimate and bootstrap confidence intervals are identical, which is a result of there not being any cases of a forecast event associated with no observed event for this threshold and particular simulation; subsequently the two statistics are identical. Naturally, the normal approximation intervals differ for this statistic because different standard errors for the statistic are used. Also in this case, the normal approximation and bootstrap intervals for hit rate (threshold=5) are very close (being identical after rounding to two decimal places as in the table), indicating that the normal approximation interval is valid despite that the qq-plots suggest otherwise.

What is clear from these simulation exercises is that in the face of iid samples, the iid bootstrap confidence intervals (using the BC_a method) are always valid, but the normal approximation intervals are only valid for statistics whose df's are approximately normal. Because the bootstrap procedure requires resampling over many iterations (and the BC_a requires an additional series of resampling), it can be slow when faced with very large samples, whereas the normal approximation intervals are quick.

Table 3 shows 95% iid bootstrap confidence intervals for some common verification scores for a particular simulation of two forecasts with observation triplet. Looking at the mean error statistic, the estimate for the simulated forecast 1 falls within the bounds for forecast 2, but the difference bounds for this statistic show that the mean error for forecast 1 is significantly different from that of the second. In this case, the estimate of mean error for the second simulated forecast does not fall within the bounds of the first, but as is seen in Figure 1, it is possible for both estimates to fall within the bounds of the other, but still have the differences be significant. The rmse for these two cases are substantially different from each other, and this is reflected by zero's being well outside the difference intervals; in this case neither estimate falls within the other's bounds. The story is similar for the rest of the intervals, and clearly, the simulated forecast 1 is superior to the simulated forecast 2 for this realization of the verification triplet.

Figure 10 shows plots of a simulated time series from an AR(2) model. Suppose the simulation represents $f - o$ for some verification set, and the data are not independent. The normal qq-plot confirms that the error simulation series are normally distributed, but the autocorrelation- and partial autocorrelation function graphs suggest that they are not independent; confirming that an AR(2) model is appropriate for the simulated data.

For simplicity, interest will be in the mean of the simulated series (i.e., the mean error), which should reasonably follow a normal df. Table 4 shows 95% confidence intervals obtained from various methods. The estimated mean (error) is about 0.03, and the estimated correlation coefficients are $\hat{\rho}_1 \approx 0.69$ and $\hat{\rho}_2 \approx -0.42$.

Figure 10: Simulation of size 1000 from an AR(2) model with $\rho_1 = 0.7$ and $\rho_2 = -0.4$. Top left is time series, top right is the normal qq-plot, and autocorrelation- and partial autocorrelation function graphs are shown in the bottom row.

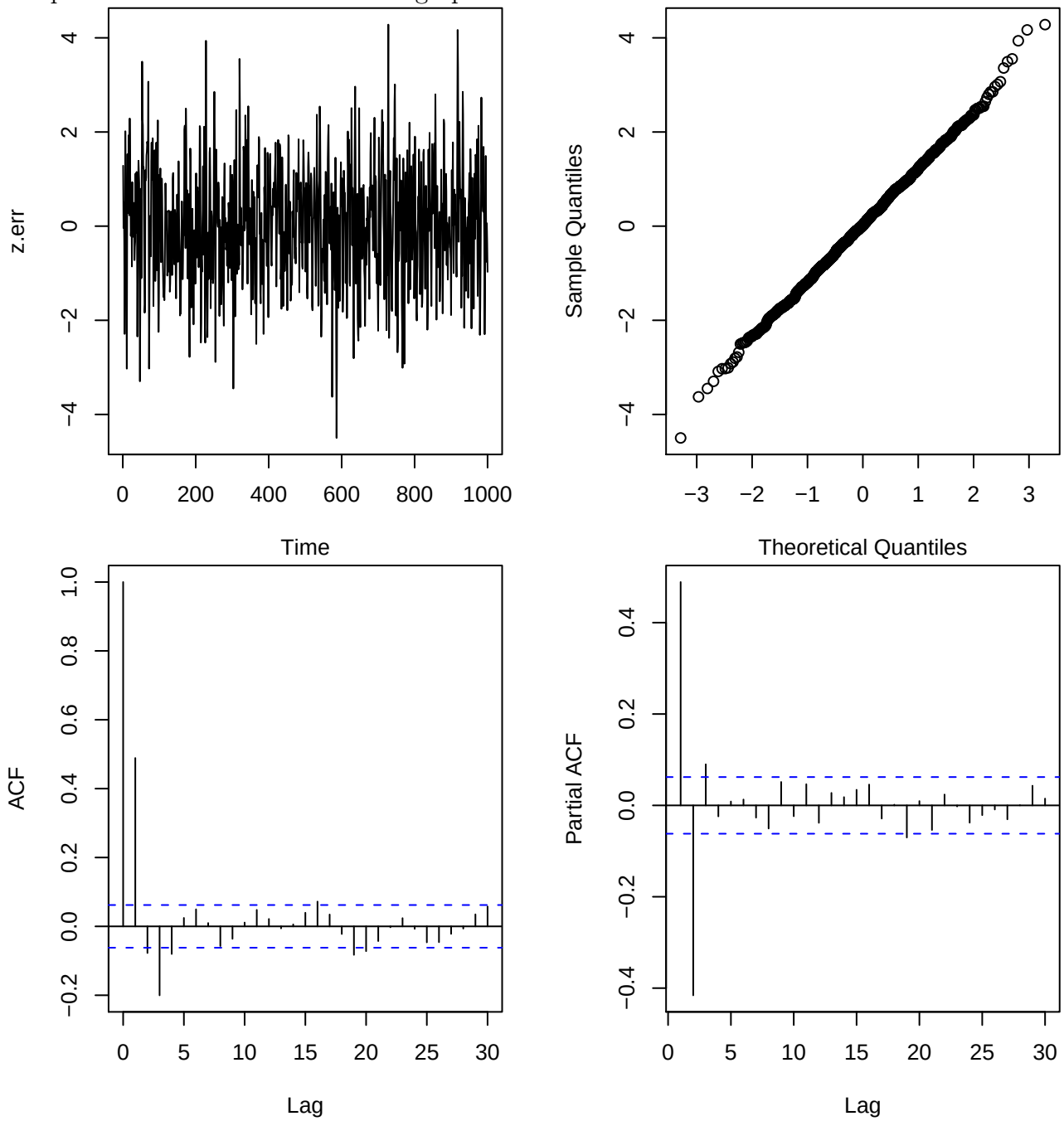


Table 3: IID Bootstrap (BC_a) 95% confidence intervals for verification statistics from two simulated forecasts against the same verification series, and for the differences between statistics for each forecast.

Statistic (Forecast 1, 2, or $1-2$) Estimate	**95% IID Bootstrap (BC_a) CI**
Mean error (1) ≈ -0.48	$(-0.54, -0.43)$
Mean error (2) ≈ -1.51	$(-1.69, -1.33)$
Mean error difference $(1-2) \approx 1.02$	(0.84, 1.22)
rmse (1) ≈ 1.03	(0.99, 1.08)
rmse (2) ≈ 3.24	(3.10, 3.39)
rmse difference $(1-2) \approx -2.21$	$(-2.36, -2.06)$
bias (1) ≈ 1.00 (threshold = 5)	(1.00, 1.01)
bias (2) ≈ 0.84 (threshold = 5)	(0.82, 0.87)
bias difference $(1-2) \approx 0.16$ (threshold = 5)	(0.13, 0.18)
bias (1) ≈ 0.41 (threshold = 14)	(0.28, 0.55)
bias (2) ≈ 2.41 (threshold = 14)	(2.03, 2.94)
bias difference $(1-2) \approx -2.00$ (threshold = 14)	$(-2.49, -1.62)$

Table 4: 95% confidence intervals for the mean (≈ 0.03) of the simulated series shown in Figure 10. Normal approximation with variance inflation factor assumes aan AR(1) model.

Normal approximation	$(-0.04, 0.11)$
IID Bootstrap (BC_a)	$(-0.04, 0.12)$
Normal approximation (with variance inflation factor)	$(-0.14, 0.21)$
Block Bootstrap ($l = \lfloor\sqrt{1000}\rfloor$)	$(-0.06, 0.12)$

Not surprisingly, the normal approximation and iid bootstrap are very close to each other. Both assume independence in the data, and because the normality assumption is valid, the two intervals should be similar. The normal approximation with variance inflation factor assumes, for simplicity, an AR(1) model for the data (see Gilleland, 2010, section 2.11 Eq. (17) and (19)), which is still not valid for this sample, but demonstrates that the confidence interval widens because effectively there are fewer sample points than 1000. The block bootstrap using the circular block bootstrap (CBB) with block length of 31 (see Gilleland, 2010, section 3.4) is also shown. This interval is wider than the iid bootstrap and normal approximation (without variance inflation factor) intervals, but narrower than the variance inflated normal approximation interval. Although a better interval might be obtained using a parametric bootstrap with an AR(2) model, the CBB interval is quite reasonable.

Acknowledgments

Support for this manuscript was provided by the Air Force Weather Agency (AFWA) (`http://www.afweather.af.mil/`), the Developmental Testbed Center (DTC) (`http://www.dtcenter.org/index.php`) and the National Center for Atmospheric Research (NCAR). The National Center for Atmospheric Research is sponsored by the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Bernadet, L., J. Wolff, L. Nance, A. Loughe, B. Weatherhead, E. Gilleland, and B. Brown, 2009: Comparison between ARW and NMM objective verification scores. *23rd Conf. on Wea. Anal. and Forec./19th Conf. on Numer. Wea. Predic., Amer. Meteor. Soc., Omaha, NE*, 1–6.

Bernardo, J. and A. Smith, 2000: *Bayesian Theory*. Wiley, New York, N.Y.

Brockwell, P. and R. Davis, 2002: *Introduction to Time Series and Forecasting*. Springer, New York., 420pp.

Coles, S., 2001: *An introduction to statistical modeling of extreme values*. Springer-Verlag, London, UK, 208 pp.

Dawid, A. and M. Stone, 1982: The functional-model basis of fiducial inference. *Ann. Statistics*, **10**, 1054–1067.

Fisher, R., 1935: The fiducial argument in statistical inference. *Ann. Eugenics*, **6**, 391–398.

Gilleland, E., 2010: Confidence intervals for forecast verification. Technical Note NCAR/TN-479+STR, NCAR.

Gilleland, E. and R. W. Katz, 2006: Analyzing seasonal to interannual extreme weather and climate variability with the extremes toolkit (extremes). *18th Conference on Climate Variability and Change, 86th American Meteorological Society (AMS) Annual Meeting, 29 Jan - 2 Feb, Atlanta, Georgia*, **P2.15**, 1–11.
URL `http://www.assessment.ucar.edu/pdf/Gilleland2006revised.pdf`

Hannig, J., H. Iyer, and P. Patterson, 2006: Fiducial generalized confidence intervals. *J. Amer. Statistical Assoc.*, **101**, 254–269, doi:10.1198/016214505000000736.

Jolliffe, I., 2007: Uncertainty and inference for verification measures. *Wea. Forecasting*, **22**, 637–650.

Jolliffe, I. and D. Stephenson, 2003: *Forecast Verification. A Practitioner's Guide in Atmospheric Science*. Wiley and Sons Ltd.

R Development Core Team, 2008: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL `http://www.R-project.org`

Wilks, D., 2006: *Statistical Methods in the Atmospheric Sciences. An Introduction*. Academic Press, San Diego, 627 pp., second edition.

Index

Bayesian
    credible intervals, 1
between-field uncertainty, 15
block bootstrap
    circular block, 24

iid noise, 7

variance inflation factor, 24

white noise, 7
within-field uncertainty, 15