

# An assessment of Bayesian bias estimator for numerical weather prediction

J. Son<sup>1</sup>, D. Hou<sup>2</sup>, and Z. Toth<sup>3</sup>

<sup>1</sup>Numerical Prediction Center KMA, Seoul, Korea

<sup>2</sup>Environmental Modeling Center/NCEP/NWS/NOAA and SAIC, Washington DC, USA

<sup>3</sup>Environmental Modeling Center/NCEP/NWS/NOAA, Washington DC, USA

Received: 24 April 2008 – Revised: 6 November 2008 – Accepted: 6 November 2008 – Published: 16 December 2008

**Abstract.** Various statistical methods are used to process operational Numerical Weather Prediction (NWP) products with the aim of reducing forecast errors and they often require sufficiently large training data sets. Generating such a hindcast data set for this purpose can be costly and a well designed algorithm should be able to reduce the required size of these data sets.

This issue is investigated with the relatively simple case of bias correction, by comparing a Bayesian algorithm of bias estimation with the conventionally used empirical method. As available forecast data sets are not large enough for a comprehensive test, synthetically generated time series representing the analysis (truth) and forecast are used to increase the sample size. Since these synthetic time series retained the statistical characteristics of the observations and operational NWP model output, the results of this study can be extended to real observation and forecasts and this is confirmed by a preliminary test with real data.

By using the climatological mean and standard deviation of the meteorological variable in consideration and the statistical relationship between the forecast and the analysis, the Bayesian bias estimator outperforms the empirical approach in terms of the accuracy of the estimated bias, and it can reduce the required size of the training sample by a factor of 3. This advantage of the Bayesian approach is due to the fact that it is less liable to the sampling error in consecutive sampling. These results suggest that a carefully designed statistical procedure may reduce the need for the costly generation of large hindcast datasets.

## 1 Introduction

Statistical methods are widely used to process Numerical Weather Prediction (NWP) products with the aim of improving the forecast. The adjustment of dynamically based (NWP) forecasts with statistical models has a long history. Model Output Statistics (MOS) techniques (e.g. Glahn and Lowry, 1972; Woodcock, 1984; Vislocky and Fritsch, 1995) have been widely used since the 1970s. It improves raw numerical forecasts by reducing model bias and filtering out the unpredictable. These statistical algorithms adjust the raw forecast based on a database of retrospective forecasts, preferably from the same model, and the corresponding observations. The size of the sample of forecast-observation pairs is crucial for the application of these algorithms. As the characteristics of the errors in NWP model output depends on the model used in generating the forecast, a large number of retrospective forecasts must be run prior to implementation of a new model or upgrading of an existing model. As NWP models are continuously improved and periodically upgraded, the cost associated with the generation of a large sample of retrospective forecasts may hinder the application of such statistical post processing algorithms. As an example, Hamill et al. (2004) suggest that full benefit of the MOS approach can be achieved with about 20 years of training data.

On the other hand, some statistical methods based on Bayes Theorem (Krzysztofowicz, 1983; Berger, 1985; Bernardo and Smith, 1994) have been developed to process NWP model products. They are able to generate probabilistic forecast from a sample of deterministic NWP model output and the corresponding truth. In contrast to the more traditional statistical approach, these Bayesian methods make use of the information from a much larger, existing sample of the truth (observation or analysis), from which the climatology



Correspondence to: J. Son  
(jhsong@kma.go.kr)

of the meteorological variable in consideration can be derived. Krzysztofowicz (1999) proposed a Bayesian Processor of Forecast (BPF) which quantifies the uncertainty in terms of probability density function of the real value of the forecast variable, given the raw forecast (NWP model output). By using the climatology distribution of the variable and a statistical relationship between the raw forecast and the verification, BPF can estimate this probability distribution function (pdf) from a relatively small sample of the forecast-truth pairs. In addition, the BPF implicitly corrects the bias in the raw forecast.

Despite its sound theoretical basis, BPF is not widely used in the statistical processing of NWP products. This paper provides a preliminary test of the simplest Bayesian algorithm, i.e., that used in bias correction.

Most Numerical Weather Prediction (NWP) products are subjected not only to random error but also to systematic error (i.e., bias). By definition, bias is the expected difference between nature and a forecast of nature. Bias arises from the limitations of the numerical models used in the integration (Toth and Pena, 2007), and their estimation and correction are of great interest in both research and forecast operations. Particularly, in the case of longer-lead time forecast, bias correction is essential for correcting model drift in the forecast. Interest in bias estimation and correction has been on the rise in recent years with the emergence of ensemble forecast products, especially those from multi-model ensembles, such as the North America Ensemble Forecast System (NAEFS). Because all forecast systems have their own systematic errors (e.g. Hou et al., 2001) and these errors would cause bias in the first and second moments of the ensemble distribution (Cui et al., 2005), they should preferably be removed before single model ensembles are combined to generate a joint, multi-model ensemble. Removing bias also improves highly quadratic scores, such as the root mean square error (RMSE).

There are various schemes of bias estimation (e.g. Déqué, 2003; Cui et al., 2005). A good bias estimation scheme has two desired characteristics. First, the estimated bias should converge to the true bias with increasing sample size used in the bias estimation (i.e., should yield an unbiased estimate of the model systematic error). Second, the rate at which the estimated bias approaches the real bias should be high. The second requirement is very important for operational applications, where the NWP model is continuously improved and periodically upgraded. Before each implementation, a retrospective data set needs to be generated to facilitate bias estimation and correction. Faster convergence of the estimated bias to its real value implies a higher quality in bias estimation, leading either to an improved forecast (with a specified size of the training sample), or reduced need for computational resources (with specified accuracy). Therefore, the rate of convergence of bias estimation is the most important issue to be analyzed when a bias correction scheme is assessed.

To assess a bias correction scheme, the use of real observation (or analysis) and forecast data sets accumulated for multiple years at operational forecast centers appears to be the most straightforward. However, such analysis/forecast data sets are typically not available. This is because every forecast model evolves as time goes by and the computer resources for regenerating retrospective forecasts are limited. To avoid this limitation, some studies used either an earlier version (cheaper for run) of an operational model (Hamill et al., 2004), or a simpler model (Gneiting et al., 2005) to generate large training data sets. Although the sample size generated in this manner is larger, it is still insufficient for a rigorous study of the basic issue, i.e. the rate of convergence in the bias estimator, which requires a specification of the climatological mean of the forecast. For this reason, a different methodology is followed in this study, by using synthetically generated time series to represent the truth and forecast. In addition to providing arbitrarily long time series to define the climate, this method rules out the regime dependence of bias, and allows a controlled and detailed analysis of the bias estimation and correction methods for better understanding of their performance.

The paper is organized as follows: Sect. 2 introduces the bias estimator corresponding to the BPF and compares it with the commonly used method of empirical bias estimation. Section 3 describes how the synthetic analysis and forecast data are generated. Results from experiments and analytical analysis with the synthetic data set are shown in Section 4, and those from a preliminary test with real NWP forecast in Sect. 5. Finally, a summary and a brief discussion of the results are presented in Sect. 6.

## 2 Empirical and Bayesian bias estimators

The mean systematic error, or bias, of a forecast system, is typically defined as the statistical expectation of the difference between forecast  $f$  and the corresponding truth  $a$ , i.e.

$$B = E(f - a) = E(f) - E(a) \quad (1)$$

and empirically estimated (e.g. Déqué, 2003; Cui et al., 2005) from a sample of size  $n$

$$\hat{B} = \frac{1}{n} \sum_{i=1}^n (f_i - a_i) \quad (2)$$

In fact, if there is no other information available except the sample of the forecast and analysis ( $a$ ,  $f$ ) pairs, this is the only way to estimate the bias. If  $n$  is small and the sample is not representative of the population,  $\hat{B}$  will be significantly different from the real bias  $B$ . In other words, to ensure a reasonable estimation of the bias, a sufficiently large sample and/or some special sampling technique is necessary.

However, in operational forecasting, the climate probability distribution function (pdf) of a meteorological variable

is often available with many years of observation or analysis. It is also a common practice in both the traditional MOS approach and the Bayesian approach (e.g. Krzysztofowicz, 1999) to assume that the forecast  $f$  is the sum of a function of the verification  $a$ , denoted by  $G(a)$ , the constant bias  $B$ , and a random error  $\varepsilon$  independent of the truth  $a$  and with zero mean, or mathematically,

$$f = G(a) + B + \varepsilon \quad (3)$$

An assumption behind Eq. (3) is that the expected value of  $G(a)$  is the same as that of the analysis  $a$  itself, i.e.

$$E[G(a)] = E(a). \quad (4)$$

By considering Eq. (4), the bias  $B$  can be expressed as the expected value of the difference between  $f$  and  $G(a)$ , i.e.

$$B = E[f - G(a)] \quad (5)$$

And can be estimated from a sample of  $(f, a)$  pairs as

$$\hat{B} = \frac{1}{n} \sum_{i=1}^n [f_i - G(a_i)] \quad (6)$$

Since  $E(\varepsilon)=0$ ,  $G(a)$  is the expected value of  $(f-B)$  with given  $a$ , i.e.

$$G(a) = E[(f - B)|a] \quad (7)$$

and it reflects the statistical relationship between the forecast and the verification. In addition, the first moment of the climate pdf of the analysis,  $E(a)$ , is used in defining  $G(a)$ . Therefore, Eq. (6) is the same as the bias estimated by BPF (Krzysztofowicz, 1999) discussed in section 1. As the Bayes Theorem is applied implicitly by using  $E(a)$ , Eq. (6) is referred as a Bayesian estimator of bias  $B$ , in contrast to the empirical bias estimator Eq. (2).

An advantage of the Bayesian Bias Estimator (BBE) over the Empirical Bias Estimator (EBE) is that its accurate form Eq. (5) holds not only for the expected value, but also for the conditional expected value, given analysis  $a$ , i.e.,

$$B = E[\{f - G(a)\}|a] \quad (8)$$

Equation (8) can be easily proved by noting that  $f-G(a)=B+\varepsilon$  and  $\varepsilon$  is a random number with 0 mean and independent of the truth  $a$ . This property of BBE indicates that the bias  $B$  can be accurately estimated by using only a subsample of  $(a, f)$  pairs with a specific value or a small range of values of analysis  $a$ , instead of a large sample spanning all of the possible values of  $a$ . Consequently, it can be used to increase the rate of convergence in bias estimation and hence reduce the required sample size if a specific accuracy is required.

The application of BBE in Eq. (6) requires specifying function  $G(a)$ . With both the traditional MOS approach and the Bayesian Forecast Processor (Krzysztofowicz, 1999), a linear relation is commonly assumed. As this assumption is

valid in most cases, it is also accepted in this study, although other functions, such as logistic function, can be used. To satisfy Eq. (4), the following linear function is the most natural choice:

$$G(a) = \alpha[a - E(a)] + E(a) \quad (9)$$

Note that this is different from the linear assumption in traditional MOS techniques in that the climatological information of the truth,  $E(a)$ , is employed. It will be noted later that this distinction is very important. With this linear function, Eq. (3) takes the form of

$$f = \alpha[a - E(a)] + B + E(a) + \varepsilon \quad (10)$$

and, with an estimation of the slope  $\alpha$  the BBE in Eq. (6) becomes

$$\hat{B} = \frac{1}{n} \sum_{i=1}^n [f_i - \hat{\alpha}(a_i - E(a))] - E(a) \quad (11)$$

With the traditional MOS approach,  $E(a)$  is unknown or not used. When the mean values of  $f$  and  $a$  are required (as in bias correction) they are often substituted with an estimation from the sample mean. It can be shown that, with this type of substitution, the BBE in Eq. (11) decays to EBE in Eq. (2). Therefore, the difference between BBE and EBE is not only in the forms of formulation, but also in the nature of the methodology. While the traditional MOS approach, including EBE, adjusts the forecast  $f$  based only on the available sample of  $(a, f)$  pairs, BBE uses some information of the truth from the whole population. When there is only a partial sample, the difference is significant. For example, when the forecast skill is very low ( $\alpha=0$ ), EBE in Eq. (2) adjusts the forecasts to the sample mean while BBE in Eq. (11) adjust them towards to the climatological mean.

For convenience in calculation and discussion, the model can be further simplified by denoting  $b=B+(1-\alpha)E(a)$  and rewriting Eqs. (10) and (11) as

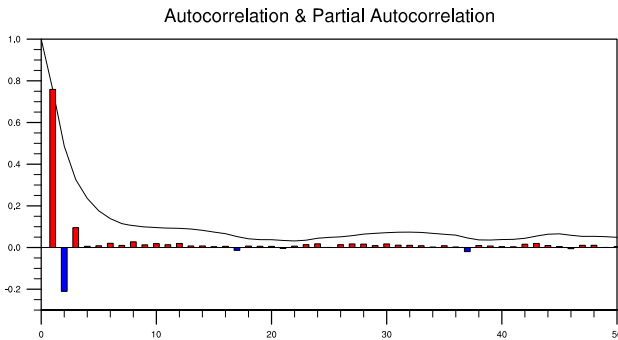
$$f = \alpha a + b + \varepsilon \quad (12)$$

and

$$\hat{b} = \frac{1}{n} \sum_{i=1}^n (f_i - \hat{\alpha} a_i) \quad (13)$$

As in the traditional MOS approach, the intercept  $b$  of the linear function in Eq. (12) is not the bias defined in Eq. (1) and a bias estimation has to be obtained by adding  $(\alpha-1)E(a)$  to the result of Eq. (13). The real simplification is by assuming  $E(a)=0$ , or working in normal space. For this special case,  $b$  is the bias defined in Eqs. (1) and (13) is the Bayesian Bias Estimator, while the Empirical Bias Estimator in Eq. (2) becomes

$$\hat{b} = \frac{1}{n} \sum_{i=1}^n (f_i - a_i) \quad (14)$$



**Fig. 1.** Autocorrelation coefficient (line) and Partial autocorrelation coefficient (histogram) as a function of time lag (days), calculated from the time series of 40 years of daily reanalysis of 2 m temperature at 37.5 N, 117.5 W.

In this article, calculation and discussion are performed in normal space with Eq. (12) as the linear model, and Eqs. (14) and (13) used as EBE and BBE, respectively. The results can be easily extended to the original space. However, when comparing the two estimators, one should keep in mind that BBE has a general form of Eq. (11) and its usage requires the information of the climatological mean of the truth.

### 3 Generation of the synthetic data set

#### 3.1 General considerations

The synthetic data set used in this study seeks to represent the truth with an arbitrarily long time series resembling the major characteristics of the observation, and express its relationship to the forecast with an analytical formula. Linear models are commonly used in the traditional MOS approach and the Bayesian processor (Krzysztofowicz, 1999), and this practice is followed in this study. As can be seen later, further simplification is made by working in a standard space and unit variance will be specified for both the truth and the forecast.

While observation is widely used to represent the truth, objective analysis (which, in addition to the observation, uses the same NWP model as used to generate the forecast) is more commonly used in the major operational centers for model verification and calibration. With this in mind, the truth used in this study is based on real analysis and referred as synthetic analysis, or simply analysis.

As this study is focused on the problem of bias correction, or the adjustment of the first moment, the adjustment of the second and higher moments of the probabilistic distribution is ignored. Therefore, the corrected forecast will have the same variance as the truth.

#### 3.2 Synthetic analysis

To represent the truth, a synthetic analysis data set was generated based on the statistics from the National Centers for Environmental Prediction (NCEP) – National Center for Atmospheric Research (NCAR) reanalysis (Kalnay et al., 1996). The reanalysis data set consists of daily analyses, from January 1959–December 1999, for a number of near-surface and upper air variables on a  $2.5 \times 2.5$  latitude/longitude global grid. Temperature at 2 m height at the grid point 37.5 N, 117.5 W (near Fresno, California) is used for this study. The selection of this point is arbitrary but comparisons with other grid points suggest that the generated time series is representative of mid-latitude regions of North America.

To focus on the basic characteristics of bias estimation methods, it is useful to disregard the fluctuations related to the annual cycle. Therefore, the reanalysis time series is standardized by subtracting climate mean from each temperature value and then divided by the standard deviation. Both the climate mean and the standard deviation are calculated from the 40-year (1959–1998) climate data for the Julian day of the year corresponding to the date under consideration.

After removing the annual cycle by standardization, the reanalysis time series is used to determine the parameters of an ARMA model, which is then applied to generate the synthetic analysis. An ARMA model (Box and Jenkins, 1976; Gershenfeld and Weigend, 1994) consists of two parts, an autoregressive (AR) part and a moving average (MA) part, and is usually referred to as an ARMA( $p, q$ ) model, where  $p$  is the order of the autoregressive part and  $q$  is the order of the moving average part. It can be written as

$$x_t = \sum_{i=1}^p \phi_i x_{t-i} + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \varepsilon_t \quad (15)$$

where  $\phi_i$  and  $\theta_i$  are the autocorrelation parameters and the moving average parameters of the model, respectively. The error terms  $\varepsilon_t$  are generally assumed to be independent and identically-distributed random variables, sampled from a normal distribution with zero mean and unit variance:  $\varepsilon_t \sim N(0, 1)$ .

To select the proper order  $p$  for the autoregression, the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) (Pourahmadi, 2001) of the normalized reanalysis time series were computed and shown in Fig. 1. It can be seen that the ACF decreases most rapidly between lags of 0 to 10 and the time series has an autocorrelation noticeably different from zero only for lags less than 20. PACF at lag  $k$  is defined as the correlation coefficient between  $X_t$  and  $X_{t+k}$  that is not accounted for by lags 1 through  $k-1$ . For an AR( $p$ ) model it drops off to zero after lag  $p$ . Therefore, it is more convenient to use PACF in identifying the order  $p$  (Quenouille, 1949). The fact that PACF vanishes after lag 3 suggests  $p=3$  is an acceptable choice. However, in order to retain as much information as possible from the

real analysis time series, a conservative selection of  $p=20$  is used. In fact, the resulted time series with  $p=3$  and  $p=20$  are very similar. The order of moving average was selected as  $q=1$ , after several tests with higher  $q$  showed no significant improvement.

The coefficients of the ARMA model in Eq. (15) (with  $p=20$  and  $q=1$ ) are estimated from the normalized reanalysis time series with a size of 14610, using subroutines in the commercial IMSL Stat/Library from Visual Numerics, Inc. The algorithm is similar to that of Box and Jenkins (1976, pages 498–500). The ARMA(20,1) model is then used to generate an arbitrarily long time series, which is used as the synthetic analysis data set. Figure 2 shows a section of the standardized reanalysis time series ( $t=1$  to 365) and a section of the synthetic analysis time series ( $t=366$  to 730), which is generated by the ARMA(20,1) model fitted to the reanalysis. The two sections are hardly distinguishable from each other. Therefore, we conclude that the synthetic analysis is a good approximation of the standardized reanalysis or observational time series. For convenience of calculation and comparison, the synthetic analysis is slightly adjusted so that its mean over a period of 100 000 days (about 270 years) is exactly 0.

Both the standardized reanalysis and the synthetic analysis exhibit variability at various frequencies. Although the annual cycle has been removed, some fluctuations with their frequency lower than the random noise still exist in the synthetic analysis time series. Consequently, the average of the time series over a period of about 100 days or shorter is significantly different from its climate mean (0). For example the periods from  $t=365$  to 465 and  $t=630$  to 730 in Fig. 2 are dominated by positive values. These lower frequency fluctuations in the synthetic analysis time series can be interpreted as quasi-seasonal variations associated with the changes in dominant circulation patterns over seasonal or longer time scales and will be further discussed in Sects. 5 and 6.

Finally, by specifying different seed value for the random number generator in running the ARMA model, a number of time series of synthetic analysis  $a$  can be generated. They are different from each other but have the same statistics and each of them is called a random case of the synthetic analysis in this study.

### 3.3 Synthetic forecasts

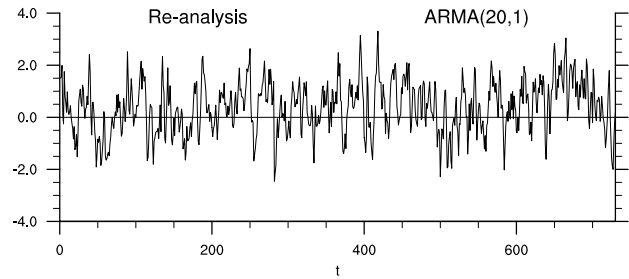
Consistent with Eq. (12), a synthetic forecast time series  $f$  is generated using the following normal-linear model:

$$f_i = \alpha a_i + \beta e_i + b \tag{16}$$

where

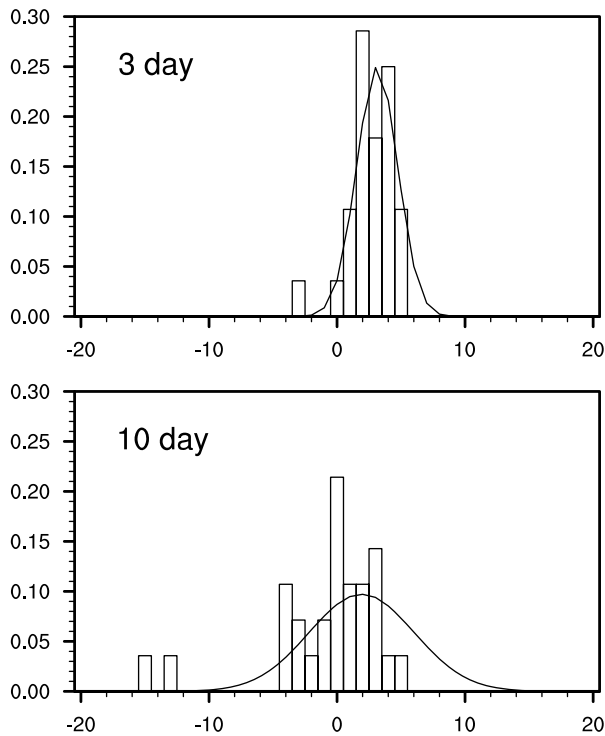
$$0 < \alpha < 1, \beta = \sqrt{1 - \alpha^2} \tag{17}$$

$e$  is a random number from a normal distribution with zero mean and unit variance, and the subscript  $i$  is the index of the time series. The choice of the relationship between  $\alpha$  and



**Fig. 2.** The time series constructed from both the real and synthetic analyses. The section  $t=1$  to 365 is from the standardized reanalysis data, and  $t=366$  to 730 from the synthetic analysis generated with ARMA(20,1).

$\beta$  in Eq. (17) is necessary for the variance of the synthetic forecast  $f$  to match that of the synthetic analysis  $a$ .  $\alpha$  can be easily shown to be the temporal correlation coefficient between the forecast  $f$  and the analysis  $a$  in standard space, or the anomaly correlation in the original space. In Krzysztofowicz (1992) this parameter is referred as Bayesian Correlation Score and shown to be meaningful for comparing alternative forecasts. Murphy and Epstein (1989) also relate it to skill scores. In this study,  $\alpha$  is varied between 1 and 0 to roughly represent NWP forecasts with lead times varying between 0 (perfect correlation, and no random error) and 15 days (no correlation and the forecast is dominated by random errors). Before proceeding to testing different bias estimation methods, we assess whether  $f$  as defined in Eq. (16) is consistent with the statistics of real NWP forecasts. In particular, we are interested to see whether the operational forecasts are approximately normally distributed with an expected value  $\alpha \bar{a}$  (the sample average of the corresponding analysis multiplied by the correlation between forecast and analysis), given analysis  $a$ . This assumption should hold for a larger sample ranging over all possible values of the analysis, and for sub samples of analysis values over a particular range. Operational Global Forecast System (GFS) forecasts of NCEP and the corresponding analysis for the period from April 2004 to August 2005 were used in the latter, more stringent test. Figure 3 shows the histograms of the sub sample of forecasts with corresponding analyses between 3.0 and 4.0 degrees, and the hypothetical normal distribution, for the 3-day and 10-day forecast. It can be seen that the normal distribution roughly fits the histogram for both cases. Chi-square (Conover, 1980; Wilks, 2006) and Kolmogorov-Smirnov (Conover, 1980) tests were performed to quantify the goodness of fit and the result of Kolmogorov-Smirnov is shown in Table 1. For 13 out of the 16 lead times, the empirical and the theoretical distribution are close to each other, justifying the use of Eq. (16).



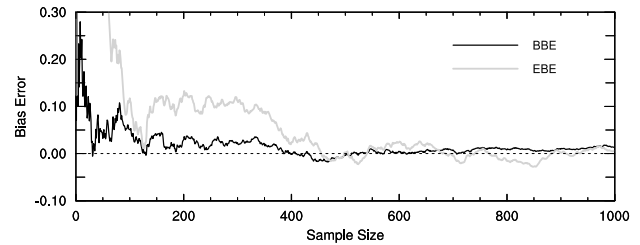
**Fig. 3.** Histogram of the real operational forecast distribution and the hypothetical pdf of the normal distribution  $N(\alpha\bar{a}, \sigma^2)$  where  $\bar{a}$  is the sample mean of the corresponding analysis,  $\alpha$  the forecast-analysis coefficient and  $\sigma^2$  the sample variance (see the text for details).

**Table 1.** The Kolmogorov-Smirnov test of the goodness of fit for various forecast lead times.  $D$  denotes maximum difference between the theoretical and the empirical cumulative distribution functions, and Prob the probability of the statistic exceeding  $D$  under the null hypothesis of equality and against the one-sided alternative. The approximation is very close for Prob less than 0.10.

Lead time	$D$	Prob	Lead time	$D$	Prob
day 1	0.2786	0.0103	day 9	0.2982	0.0053
day 2	0.1656	0.1934	day 10	0.2944	0.0060
day 3	0.1867	0.1253	day 11	0.3613	0.0004
day 4	0.4537	0.0000	day 12	0.2766	0.0110
day 5	0.3709	0.0003	day 13	0.3145	0.0029
day 6	0.4311	0.0000	day 14	0.2860	0.0080
day 7	0.3959	0.0001	day 15	0.2557	0.0211
day 8	0.2668	0.0150	day 16	0.1246	0.3871

#### 4 Results with the synthetic data set

With the synthetic data set described in Sect. 3, the Bayesian Bias Estimator is Eq. (13) and, if using the specified value of the correlation coefficient, it takes the form of



**Fig. 4.** Bias error,  $\hat{b}-b$ , as a function of sample size  $n$ , for a randomly selected case using the EBE (grey) and BBE (black).  $\alpha=0.3$ .

$$\hat{b} = \frac{1}{n} \sum_{i=1}^n (f_i - \alpha a_i) \tag{18}$$

In this section, BBE is compared with EBE defined in Eq. (14), in terms of accuracy of the bias estimate and requirement of sample size for a specified accuracy. For each case of the analysis time series, a forecast time series is generated with a specified  $\alpha$  using Eq. (16) and the bias estimation  $\hat{b}$  is calculated from the first  $n$  ( $a, f$ ) pairs for  $n=1, 2, 3, \dots$  and so on. This sequential sampling, without skipping, is commonly used in both research and operations of NWP output processing (e.g. Hamill et al., 2004; Cui et al., 2005).

$b=1$  is assumed in the calculations but the results can apply to any value of  $b$  because the bias error is independent of the bias level. This can be shown mathematically. For the synthetic forecast used in this study (Eq. 16), the two bias estimators can be generalized by

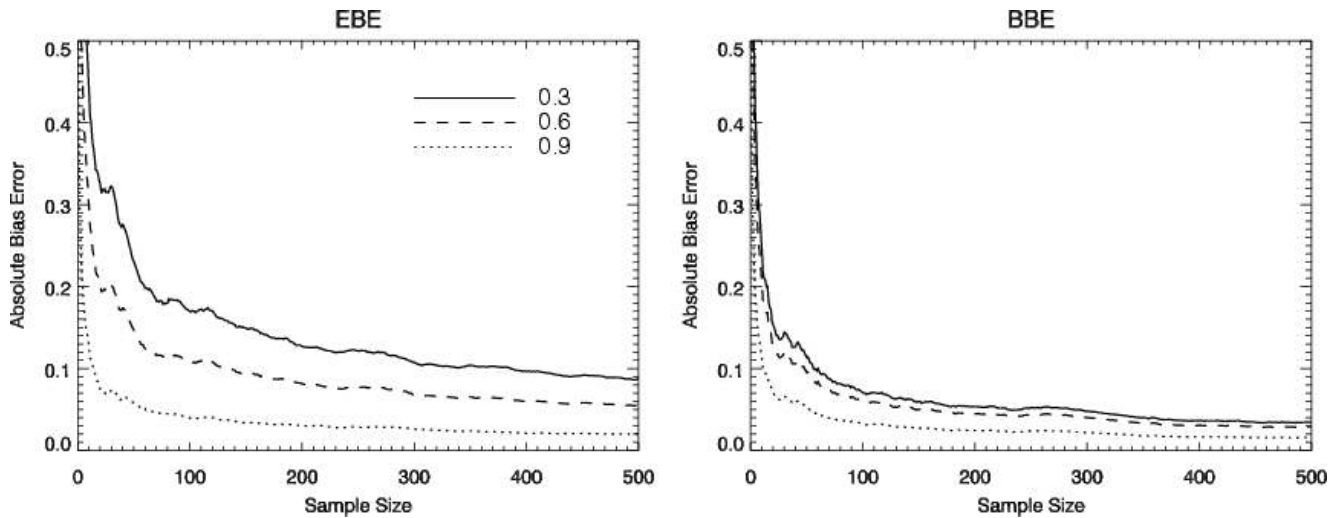
$$\hat{b}_n = \frac{1}{n} \sum_{i=1}^n [(\alpha - \gamma)a_i + \beta e_i + b] \tag{19}$$

where  $\gamma=1$  for EBE and  $\gamma=\alpha$  for BBE. From Eq. (19) one can see that

$$\begin{aligned} \hat{b} - b &= \frac{1}{n} \sum_{i=1}^n [(\alpha - \gamma)a_i + \beta e_i + b] - b \\ &= \frac{1}{n} \sum_{i=1}^n [(\alpha - \gamma)a_i + \beta e_i] + \frac{1}{n} \sum_{i=1}^n b - b \\ &= \frac{1}{n} \sum_{i=1}^n [(\alpha - \gamma)a_i + \beta e_i] \end{aligned} \tag{20}$$

is independent of  $b$ .

Figure 4 depicts the error in the estimated bias,  $b-\hat{b}$  as a function of the sample size  $n$  in a randomly selected case with  $\alpha=0.3$  (corresponding to lead time of 12 days). The error is characterized by large values and rapid variations with  $n<100$ . However, compared with EBE, the error with BBE is much smaller. For  $100<n<400$ , the error becomes relatively stable and its size is significantly reduced with both methods, but BBE clearly outperforms EBE with a ratio of the error size of about 1/5. When  $n$  exceeds 400, the error becomes even smaller and the difference between the two methods becomes less distinctive except that oscillations are still visible with the traditional approach EBE.

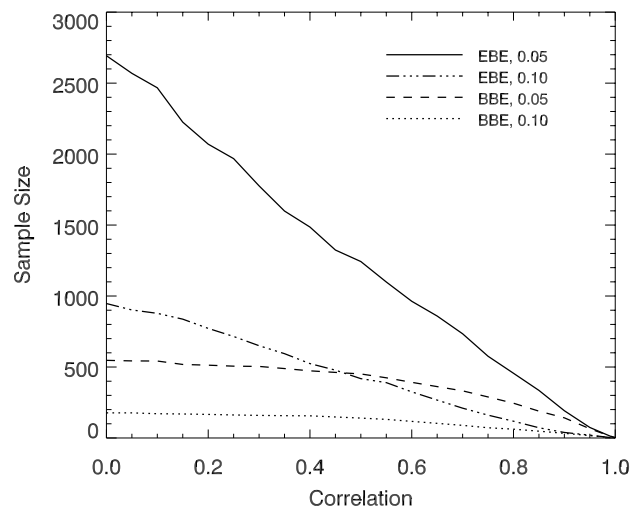


**Fig. 5.** Absolute bias error of EBE (left) and BBE (right), averaged over 100 random cases, as functions of sample size  $n$ . The solid, dashed and dotted lines correspond to the correlation between forecast and analysis of 0.3, 0.6 and 0.9, respectively.

The performance of the two bias estimators is measured by the absolute bias error, or the absolute value of the bias error, averaged over 100 randomly selected cases. This measure, as plotted in Fig. 5, smoothing out the noises and reducing fluctuations seen in individual cases, represents the general behavior of the two bias estimators. As the sample size increases from 1 to the order of 100, the absolute error in the bias estimate decreases steadily with minor fluctuations in both methods, indicating that  $\hat{b}$  gradually converges to the actual bias  $b$ . Generally speaking, the rate of convergence is higher with BBE than that with EBE. After  $n=200$ , the absolute bias error continues to decrease steadily and converges to zero as the sample size increases, with a significantly lower error level with the Bayesian approach. Comparing the three curves corresponding to various correlation values in each panel of Fig. 5, it can be seen that the error is larger for forecasts with lower correlation. Another comparison between the two panels suggests that the biggest impact of the Bayesian approach is on these less skillful forecasts.

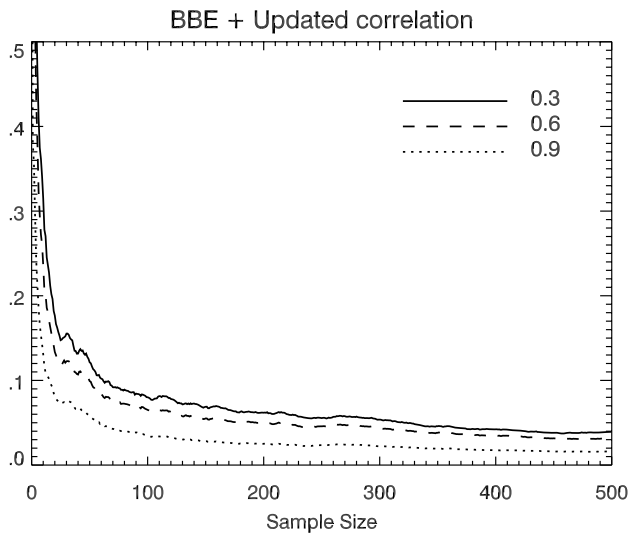
From Eq. (20), it can be seen that the larger error in the estimated bias, its profound fluctuations and slower convergence toward the real bias level are caused by the analysis or the truth, which appears in the summation for EBE, but not in BBE. As discussed in Sect. 3.2, there is noticeable lower frequency fluctuations associated with the analysis time series. At the existence of these quasi-seasonal variations, the sequential sampling is likely to result in a sample substantially different from the population of the  $(f, a)$  pairs, if the sample size  $n$  is about 100 or smaller.

A major argument for a large sample size in NWP output processing (including bias correction) is the higher accuracy achieved in the bias estimation. Figure 6 shows the sample size required for the absolute error to be less than specific



**Fig. 6.** The sample size required for the absolute bias error to be less than 0.05 (solid line for EBE and dashed line for BBE) and 0.1 (dot-dashed line for EBE and the dotted line for BBE) as a function of correlation  $\alpha$ .

thresholds of 0.05 and 0.1. With EBE, the required sample sizes are 2500 days and 850 days respectively when the correlation is 0.1. If BBE is used, they are only 550 days and 180 days. In other words, the Bayesian method can reduce the required training sample size by a factor of 4 to 5, indicating a significant reduction in computational expenses in generating the hindcast data set. For forecast of moderate skill, such as  $\alpha=0.4$ , the factor of reduction is about 3. As the correlation  $\alpha$  increases to 1.0 the two approaches are virtually the same, and this is clear from Eqs. (19) and (20). Another



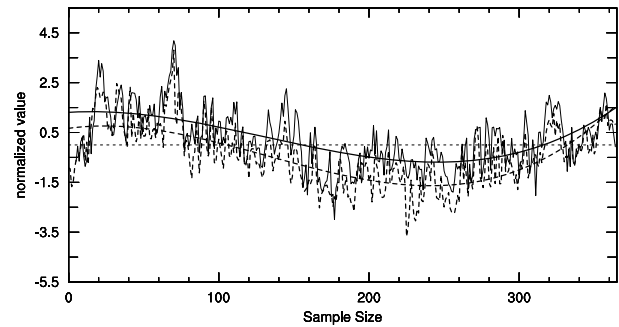
**Fig. 7.** The time series of the absolute bias error of BBE, averaged over 100 cases. Unlike in the right panel of Fig. 5, the calculation is based on the estimated value of forecast-analysis correlation from the training sample, instead of the exact value specified in generating the forecast data set.

advantage of the Bayesian Bias Estimator is that the required sample size is less dependent on the level of the forecast correlation. As seen from Fig. 6, when the correlation is less than 0.6, nearly identical sample sizes can be used to meet the required accuracy, regardless of the skill level.

In the above discussions, it is assumed that the parameter  $\alpha$  is known. If the Bayesian Bias Estimator is to be applied in operational forecast, the parameter has to be estimated from the available sample of analysis-forecast pairs. As noted in Sect. 3.3,  $\alpha$  is the correlation coefficient between the two time series, the forecast  $f$  and the analysis  $a$ . The accuracy of its estimation is also dependent on the sample size  $n$ . Calculations show that the estimated correlation coefficient approaches the real value of  $\alpha$  as  $n$  increases (not shown) and the difference is small with  $n > 100$ . This suggests that the rate of convergence is similar to that of the bias estimation and thus will not hinder the application of BBE. Bias estimation with BBE is re-conducted by using this estimated  $\alpha$  to replace the specified value, and the result is shown in Fig. 7. Comparison of Fig. 7 and the right panel of Fig. 5 reveals that the value of the absolute error, and its change with increasing  $n$ , are very similar in the two calculations, and the error is only slightly larger when the correlation is estimated from the training data set. The difference is negligible when the training sample size is larger than 100.

## 5 Preliminary results with real forecast

A rigorous test of the bias estimators BBE and EBE with a real NWP forecast data set requires a very long time series



**Fig. 8.** The time series of 120 h forecast of temperature at 2 m height from GEFS control forecast (solid line), and the corresponding analysis (dashed line) at 37.5 N, 117.5 W (near Fresno, California). The smooth curves are the moving average (with a window of 45 days, see the text) of the corresponding time series.

of the model output from the same model. Frequent model upgrading and improvements made the data accumulated at operational NWP centers over the last a few decades unsuitable for such a test. Even the reforecast data set generated using a much older model version (Hamill et al., 2004) can not match the length of the analysis data set used in this study to define the climate mean. Nevertheless, a preliminary test was conducted and the result is presented in this section. The forecast data set used is the 120 h forecast of temperature at 2 m height, output from Global Ensemble Forecast System (GEFS) running operationally at NCEP, during the year of 2005. All 11 member forecasts and the corresponding verifying analysis are standardized using the mean and standard deviation calculated from the 40-year climate data, as described in Sect. 3.2. The forecast model is NCEP's Global Forecast System (GFS) at T126 horizontal resolution with 28 levels in the vertical.

The time series of the analysis (dashed) and the control forecast (solid) are depicted in Fig. 8 with a cubic polynomial expression (which is equivalent to a moving average with a window of about 45 days) plotted as smooth curves. For the analysis, the following can be observed: (1) within this 360-point sample, there are more negative values than positive; (2) the period day 1–100 is dominated by positive values, and day 100–320 by negative values; (3) the sample mean of the analysis is negative. In other words, this one year sample is not representative of the climatology (population) in terms of its mean and distribution due to the existence of quasi seasonal variations. This is the same as what is seen in Fig. 2 for the real reanalysis and the synthetic analysis. From the moving averages of the two time series, represented by the smooth curves, it can be seen that the forecast is larger than the analysis on most days and the difference is generally larger when the analysis is negative. This suggests that the bias  $b$  is positive and the linear-normal relationship between forecast and analysis, i.e. Eq. (16), holds for this set



of real forecasts. Based on the results with the synthetic data discussed in Sect. 4, one expects that BBE will be a better bias estimator than EBE.

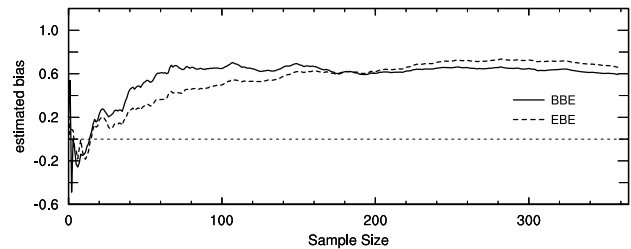
Figure 9 plots the bias estimation from BBE (solid) and EBE (dashed). It can be seen that the BBE estimation becomes stable earlier, reaching 0.8 at day 50, and the variation is smaller afterwards. In contrast, the EBE estimation reaches the same 0.8 level at day 180 and the variation after that is also larger. The properties of the two curves are very similar to what are seen in Fig. 4 for the synthetic data. Based on these observations, one has reasons to believe that 0.8 is a reasonable estimate of the real bias, which cannot be exactly determined from the unrepresentative partial sample. To reach this optimal estimation of 0.8, BBE requires a significantly smaller sample size than EBE does. While believing that these observations are valid, we must admit that a more rigorous investigation is needed to confirm the conclusion.

## 6 Discussion and conclusions

There are two reasons to work with synthetic analysis and forecast data sets in the investigation of the performance of various bias correction methods and other statistical processing schemes. First, it is possible to generate a time series as long as a user wants, without too much computational expense. Second, employment of synthetic data sets makes it possible to perform analytical analysis so the results of the experiments can be compared with theoretical solutions to thoroughly understand the performance of different schemes.

The synthetic analysis time series used in this study are generated based on the 2 m temperature reanalysis and retain the major characteristics in its temporal variation. The synthetic forecast is generated by a linear-normal model which reflects the forecast-analysis correlation, random error and the systematic bias in the real forecasts. The linear-normal model, although not the only choice, is the simplest statistical descriptions of the relationship between the forecast and the analysis. In addition, this model has been widely used since the MOS technique was proposed. Therefore, the results of this study are based on a realistic data set and the results can be applicable in real cases. However, the relationship between the forecast and the analysis in real forecasting may not follow Eq. (3) as strictly as in this synthetic case, and the parameters in the analytic form of  $G(a)$  may not be estimated as accurately as the correlation coefficient  $\alpha$ . Therefore, the advantage of the Bayesian approach is expected to be less impressive than what is seen in this study.

Two bias estimators are compared in this study. Using information from the climatological distribution of the meteorological variable in consideration, and the statistical relationship between the forecast and the verification (analysis) inferred from the available sample of limited size, the Bayesian Bias Estimator (BBE) overperforms the tradition-



**Fig. 9.** The estimated bias from BBE(solid line) and EBE(dashed line) from the data in Fig. 8.

ally used empirical approach (EBE). The formulation of BBE is effectively independent of the value of the analysis and thus requires a smaller-sized training sample for a prescribed accuracy in the estimated bias. This is important in reducing computational cost in operational NWP product processing, as the numerical weather prediction models are upgraded frequently and the initial training sample at each implementation is provided by reforecast. While Hamill et al. (2004) demonstrated that the ensemble reforecast of large sample size using an older model has significant value in improving medium range forecast skill, the difference in the error characteristics between the current and the older model is a potential problem. If a reforecast data set is to be generated each time the model is upgraded, the extra computational cost required to update the large archive of hind casts may limit the application of the method. On the other hand, using the latest model and a shorter archive is also competitive in terms of bias correction (Cui et al., 2005). With the reduced requirement for sample size with the Bayesian approach, maximum benefit can be achieved with minimum cost in running the reforecast with the latest model, only for a short period. As a rule in operational forecast centers, an experimental run with the new model has to be executed in parallel to the operational run with the older model for at least a few months before the official implementation. Therefore, a sample large enough with the Bayesian approach may already exist by the time the new model is implemented for operation. In this case, the extra expense can be eliminated.

It should be pointed out that the requirement for a larger sample size by the traditional approach EBE is largely from the consecutive sampling and the existence of the quasi-seasonal variations in the meteorological variable in consideration. As discussed in Sect. 3.2, these fluctuations with lower frequencies in the synthetic time series, to some extent, reflects the characteristics of the real reanalysis time series. Although it is not clear how well they reflect the nature, similar characteristics are found with the real analysis in a time period not covered by the 40-year reanalysis data set used to generate the synthetic analysis. This suggests these fluctuations with lower frequencies do exist in the real world. The required sample size will be reduced with EBE if the population is sampled randomly, instead of in a consecutive manner

as in the present study. Hamill et al. (2004) showed that the performance score can be improved when the same number of  $(a, f)$  data pairs is used but one pair is selected every 2, 3, 4 or 5 days. Therefore, the requirement of a larger sample size in EBE and other traditional algorithms is actually a requirement for the representativeness of the training sample. In the current study, the seasonality in the synthetic analysis time series makes a consecutive sample less representative of the population and thus the traditional approach (EBE) leads to an estimate of conditional bias (given analysis) instead of the overall bias. This problem can be avoided by using the Bayesian Bias Estimator (BBE) as its calculation is less affected by the analysis or observation.

In summary, Bayesian approach can improve bias estimation by using the climatology and the forecast-observation relationship. The same approach may be extended to other methods of NWP output processing, including the traditional MOS type techniques and adjustment of higher moments, although further investigation is required to further address these issues.

*Acknowledgements.* Much of the computation and analysis leading to this paper was performed during the first author's visit to Environmental Modeling Center (EMC)/NCEP/NOAA. Dr. Stephen Lord, the director of EMC is acknowledged for his support of the visit and this research. The authors are grateful to Roman Krzysztofowicz for discussions that substantially improved the manuscript, Yuejian Zhu for providing reanalysis and operational forecast data, and Bo Cui for the help with the computation. We thank Vladimir Krasnopolsky and George Trojan for reviewing the manuscript with constructive suggestions, and Mary Hart for the help with improving the English of the manuscript. Two Anonymous reviewers are acknowledged for their comments and suggestions that led to further improvement of the revised version of the manuscript.

Edited by: O. Talagrand

Reviewed by: two anonymous referees

## References

- Berger, J. O.: Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York, 1985.
- Bernardo, J. M. and Smith, A. F. M.: Bayesian theory. Wiley, New York, 1994.
- Box, G. E. P. and Jenkins, F. M.: Time Series Analysis: Forecasting and Control, 2nd Ed. Holden-Day, Oakland, CA, 1976.
- Conover, W. J.: Practical Nonparametric Statistics, 2nd Edition, John Wiley & Sons, New York, 1980.
- Cui, B., Toth, Z., Zhu, Y., Hou, D., and Beaugard, S.: Statistical post-processing of operational and CDC hindcast ensembles, Preprints, 21st Conference on weather Analysis and Forecasting/17th Conference on Numerical Weather Prediction. Washington DC, 1–5 Aug 2005, Amer. Meteor. Soc., 12B.2, 2005.
- Déqué, M.: Continuous Variables. Forecast Verification: A Practitioner's Guide in Atmospheric Science, edited by: Jolliffe, I. T. and Stephenson, D. B., John Wiley & Sons, Ltd, 2003.
- Glahn, H. R. and Lowry, D. A.: The use of model output statistics (MOS) in Objective weather forecasting, J. Appl. Meteor., 11, 1202–1211, 1972.
- Gershenfeld, N. A. and Weigend, A. S.: The future of Time Series: Learning and Understanding. Time Series Prediction, edited by: Weigend, A. S. and Gershenfeld, N. A., Addison-Wesley Publishing Company, Reading, MA., 1–70, 1994.
- Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T.: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation, Mon. Weather Rev., 133, 1098–1118, 2005.
- Hamill, T. M., Whitaker, J., and Wei, X.: Ensemble Reforecasting: Improving Medium-Range Forecast Skill Using Retrospective Forecast, Mon. Weather Rev., 132, 1434–1447, 2004.
- Hou, D., Kalnay, E., and Drorgemeier, K. K.: Objective verification of the SAMEX'98 ensemble forecasts, Mon. Weather Rev., 129, 73–91, 2001.
- Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, B., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Jenne, R., and Joseph, D.: The NCEP/NCAR 40-Year Reanalysis Project, B. Am. Meteor. Soc., 77, 437–471, 1996.
- Krzysztofowicz, R.: Why Should a Forecaster and a Decision Maker Use Bayes Theorem, Water Resour. Res. 19, 327–336, 1983.
- Krzysztofowicz, R.: Bayesian correlation Score: A utilitarian measure of forecast skill, Mon. Weather Rev., 19, 208–219, 1992.
- Krzysztofowicz, R.: Bayesian forecasting via deterministic models. Risk Analysis, 19, 739–749, 1999.
- Murphy, A. H. and Epstein, E. S.: Skill Scores and Correlation Coefficients in Model Verification, Mon. Weather Rev., 117, 572–581, 1989.
- Pourahmadi, M.: Foundations of time series analysis and prediction theory. Wiley-Interscience, 448 pp., 2001.
- Toth, Z. and Pena, M.: Data assimilation and numerical forecasting with imperfect models: The mapping paradigm, Physics D., 230, 146–158, 2007.
- Quenouille, M.: Approximation tests of correlation in time series, J. R. Statist. Soc. B., 11, 18–84, 1949.
- Vislocky, R. L. and Fritsch, J. M.: Improved model output statistics forecast through model consensus, B. Am. Meteor. Soc., 76, 1157–1164, 1995.
- Wilks, D. S.: Statistical Methods in the Atmospheric Sciences: An Introduction, Academic Press, 2006.
- Woodcock, F.: Australian Experimental Model Output Statistics Forecast of Daily Maximum and Minimum Temperature, Mon. Weather Rev., 112, 2112–2121, 1984.