# NEW UNBIASED SYMMETRIC METRICS FOR EVALUATION OF THE AIR QUALITY MODEL

Shaocai Yu*, Brian Eder*++, Robin Dennis*++, Shao-hang Chu**, Stephen Schwartz***
*Atmospheric Sciences Modeling Division
National Exposure Research Laboratory,
** Office of Air Quality Planning and Standards,
U.S. EPA, NC 27711
*** Atmospheric Sciences Division,
Brookhaven National Laboratory, Upton, NY 11973
e-mail: **yu.shaocai@epa.gov**
Voice (919) 541-0362 Fax (919) 541-137

## 1. INTRODUCTION

Although the operational evaluations for different air quality models have been intensively performed for regulatory purposes in the past years, the resulting array of statistical metrics are so diverse and numerous that it is difficult to judge the overall performance of the models. Some statistical metrics can cause misleading conclusions about the model performance. In this paper, a new set of unbiased symmetric metrics for the operational evaluation is proposed and applied in real evaluation cases.

## 2.0 QUANTITATIVE METRICS RELATED TO THE OPERATIONAL EVALUATION AND THEIR EXAMINATIONS

There are a lot of debates on how to present the relative differences between the model and observations. The traditional metrics (such as mean normalized bias ($M_{NB}$), mean normalized gross error ($M_{NGE}$), normalized mean bias ($N_{MB}$) and normalized mean error ($N_{ME}$), see Table 1) used in past model performance evaluations have generally used the observations to normalize the bias and error. There are two problems that may mislead conclusions with this approach, i.e., (1) the values of $M_{NB}$ and $N_{MB}$ can grow disproportionately for overpredictions and underpredicitons because both values of $M_{NB}$ and $N_{MB}$ are bounded by –100% for underprediction; (2) the values of $M_{NB}$ and $M_{NGE}$ can be significantly influenced by some points with trivially low values of observations (denomination).

++ On assignment from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce

In this study, we propose new metrics to solve the symmetrical problem between overprediction and underprediction following the concept of factor. Theoretically, factor is defined as ratio of model prediction to observation if the model prediction is higher than the observation, whereas it is defined as ratio of observation to model prediction if the observation is higher than the model prediction. Following this concept, the mean normalized factor bias ($M_{NFB}$), mean normalized gross factor error ($M_{NGFE}$), normalized mean bias factor ($N_{MBF}$) and normalized mean error factor ($N_{MEF}$) are proposed and defined as follows:

$$M_{NFB} = \frac{1}{N}\sum_{i=1}^{N} F_i ,$$

where $F_i = (\frac{M_i}{O_i} - 1.0)$ if $M_i \geq O_i$,

$$F_i = (1.0 - \frac{O_i}{M_i}) \text{ if } M_i < O_i \qquad (1)$$

$$M_{NGFE} = \frac{1}{N}\sum_{i=1}^{N} |F_i| ,$$

where $F_i = (\frac{M_i}{O_i} - 1.0)$ if $M_i \geq O_i$,

$$F_i = (1.0 - \frac{O_i}{M_i}) \text{ if } M_i < O_i \qquad (2)$$

If $\overline{M} \geq \overline{O}$

$$N_{MBF} = (\frac{\sum_{i=1}^{N} M_i}{\sum_{i=1}^{N} O_i} - 1) = \frac{\sum_{i=1}^{N}(M_i - O_i)}{\sum_{i=1}^{N} O_i} = (\frac{\overline{M}}{\overline{O}} - 1)$$

If $\overline{M} < \overline{O}$

$$N_{MBF} = (1 - \frac{\sum_{i=1}^{N} O_i}{\sum_{i=1}^{N} M_i}) = \frac{\sum_{i=1}^{N}(M_i - O_i)}{\sum_{i=1}^{N} M_i} = (1 - \frac{\overline{O}}{\overline{M}}) \quad (3)$$

$$N_{MEF} = \frac{\sum_{i=1}^{N}|M_i - O_i|}{\sum_{i=1}^{N} O_i} = \frac{M_{AGE}}{\overline{O}}, \text{ if } \overline{M} \geq \overline{O},$$

$$= \frac{\sum_{i=1}^{N}|M_i - O_i|}{\sum_{i=1}^{N} M_i} = \frac{M_{AGE}}{\overline{M}}, \text{ if } \overline{M} < \overline{O}, \quad (4)$$

where $M_i$ and $O_i$ are values of model (prediction) and observation at time and/or location *i*, respectively, N is number of samples (by time and/or location). The values of $M_{NFB}$, and $N_{MBF}$ are linear and not bounded (range from -∞ to +∞). Like $M_{NB}$ and $M_{NGE}$, $M_{NFB}$ and $M_{NGFE}$ can have another general problem when some observation values (denomination) are trivially low and they can significantly influence the values of those metrics. $N_{MBF}$ and $N_{MEF}$ can avoid this problem because the sum of the observations is used to normalize the bias and error,. The above formulas of $N_{MBF}$ and $N_{MEF}$ can be rewritten for $\overline{M} \geq \overline{O}$ case as follows:

$$N_{MBF} = \frac{\sum_{i=1}^{N} M_i}{\sum_{i=1}^{N} O_i} - 1 = \frac{\sum_{i=1}^{N}(M_i - O_i)}{\sum_{i=1}^{N} O_i}$$

$$= \sum_{i=1}^{N}[\frac{O_i}{\sum_{i=1}^{N} O_i}\frac{(M_i - O_i)}{O_i}] \quad (5)$$

$$N_{MEF} = \frac{\sum_{i=1}^{N}|M_i - O_i|}{\sum_{i=1}^{N} O_i}$$

$$= \sum_{i=1}^{N}[\frac{O_i}{\sum_{i=1}^{N} O_i}\frac{|M_i - O_i|}{O_i}] \quad (6)$$

The above two equations show that $N_{MBF}$ and $N_{MEF}$ are actually the results of summaries of normalized bias ($M_{NB}$) and error ($M_{NGE}$) with the observational concentrations as a weighting function, respectively. $N_{MBF}$ and $N_{MEF}$ have both advantages of avoiding dominance by the low values of observations in normalization like $N_{MB}$ and $N_{ME}$ and maintaining adequate evaluation symmetry like fractional bias ($F_B$) and fractional gross error ($F_{GE}$) (see Table 1). Although $F_B$ and $F_{GE}$ can solve the symmetrical problem between overprediction and underprediction, what the metrics $F_B$ and $F_{GE}$ measure is not clear because the model prediction is not evaluated against observation but average of observation and model prediction. In addition, the scales of $F_B$ and $F_{GE}$ are not linear and are seriously compressed beyond ±1 as $F_B$ and $F_{GE}$ are bounded by ±2 and +2, respectively. The meanings of $N_{MBF}$ and $N_{MEF}$ are also very clear. The meanings of $N_{MBF}$ can be interpreted as follows: if $N_{MBF} \geq 0$, for example, $N_{MBF}$ =1.2, this means that the model overpredicts the observation by a factor of 2.2 (i.e., $N_{MBF}$+1=1.2+1=2.2); if $N_{MBF}$ <0, for example, $N_{MBF}$ =-0.2, this means that the model underpredicts the observation by a factor of 1.2 (i.e., $N_{MBF}$-1=-0.2-1=-1.2).

To test the reliabilities of other quantitative metrics listed in Table 1 and newly proposed metrics, a dataset for a real case of model and observation for aerosol $NO_3^-$ was separated into four regions as shown in Figure 1, i.e., region 1 for model/observation<0.5, region 2 for 0.5≤model/observation≤1.0, region 3 for 1.0<model/observation≤2.0 and region 4 for 2.0<model/observation. Then, each metric in Table 1 was applied to different combinations of data in each region of Figure 1. As shown in Table 2, for the only data in region 1 with model/observation<0.5, i.e., the model underpredicted all observations by more than a factor of 2, $M_{NB}$, $N_{MB}$, $F_B$, $N_{MFB}$, $M_{NFB}$ and $N_{MBF}$ are –0.82, -0.78, -1.43, -1.28, -36.67, and –3.58, respectively. Obviously, only normalized mean bias factor ($N_{MBF}$) gives reasonable description of model performance, i.e., the model underpredicted the observations by a factor of 4.58 in this case. For the only data in region 4 with model/observation>2 (combination 4 in Table 2), $M_{NB}$, $N_{MB}$, $F_B$, $N_{MFB}$, $M_{NFB}$ and $N_{MBF}$ are 4.27, 2.25, 1.12, 1.06, 4.27 and 2.25, respectively. The results of $N_{MBF}$ and $N_{MB}$ reasonably indicate that the model overpredicted the observations by a factor of 3.25.

For the results of each metrics on combination case of regions 1 and 4 data in Figure 1 (i.e., 1+4 case in Table 2). $M_{NB}$, $N_{MB}$, $F_B$, $N_{MFB}$, $M_{NFB}$ and

$N_{MBF}$ are 1.50, 0.06, -0.27, 0.06, -18.02 and 0.06, respectively. Both $N_{MB}$ and $N_{MBF}$ show that the model slightly overpredicted the observations by a factor of 1.06, while $F_B$ (-0.27) shows that the model underpredicted the observations. This shows that the value of $F_B$ can sometimes result in a misleading conclusion as well. This specific case shows that it is not wise to use $F_B$ as an evaluation metric. Although the model mean (1.54 µg $m^{-3}$) is close to that of observation (1.45 µg $m^{-3}$), both $N_{ME}$ and $N_{MEF}$ (both of them are equal to 1.19 in Table 2) show that gross error between observations and model results is 1.19 times of mean observation. The calculation results of combination case 1+4 indicate that the good model performance can be concluded only under the condition that both relative bias ($N_{MBF}$) and relative gross error ($N_{MEF}$) meet the certain performance standards (or criteria). For the all data in Figure 1 (combination case 1+2+3+4 in Table 2), $M_{NB}$, $N_{MB}$, $F_B$, $N_{MFB}$, $M_{NFB}$ and $N_{MBF}$ are 0.96, 0.09, -0.13, 0.09, -10.75 and 0.09, respectively. Both $N_{MB}$ and $N_{MBF}$ show that the mean model only overpredicted the mean observation by a factor of 1.09. However, the gross error ($N_{MGE}$) between the model and observation is 0.77 times as high as observation. The scatter plot of Figure 1 also shows the large scatter between model and observation.

On the basis of the above analyses and test, it can be concluded that our proposed new statistical metrics (i.e., $N_{MBF}$ and $N_{MEF}$) on the basis of concept of factor can show the model performance reasonably with advantages of both avoidance of domination by the low values of observations and symmetry. These new metrics use observational data as only reference for the model evaluation and their meanings are also very clear and easy to explain.

## 3.0 APPLICATIONS OF NEW METRICS OVER THE US

The newly proposed metrics have been applied to evaluate performance of the US EPA Models-3/Community Mutiscale Air Quality (CMAQ) model system on $PM_{2.5}$ $SO_4^{2-}$ and $NO_3^-$ over the US. The test periods are from June 15 to July 17, 1999 and January 8 to February 18, 2002. As shown in Table 3, both $N_{MBF}$ (0.03 to 0.08) and $N_{MEF}$ (0.24 to 0.27) for weekly data of $SO_4^{2-}$ from CASTNet are lower than those of 24-hour data from IMPROVE, SEARCH and STN ($N_{MBF}$ =-0.19 to 0.22, and $N_{MEF}$ =0.42 to 0.46). For $PM_{2.5}$ $NO_3^-$, both $N_{MBF}$ (-0.96 to 0.59) and $N_{MEF}$ (0.80 to 1.70) for SEARCH, CASTNet, and IMPROVE data in 1999 and 2002 are larger. Figure 3 shows that there are large scatter between modeled and observed $NO_3^-$. More efforts in model development for simulating aerosol $NO_3^-$ are needed in future.

Table1. Summary of traditional metrics

| Metrics | Mathematical Expression |
|---|---|
| **(1) Mean** | |
| Correlation coefficient | $r = \frac{\sum_{i=1}^{N}(M_i - \overline{M})(O_i - \overline{O})}{\{\sum_{i=1}^{N}(M_i - \overline{M})^2 \sum_{i=1}^{N}(O_i - \overline{O})^2\}^{\frac{1}{2}}}$ |
| **(2) Difference** | |
| Mean Bias | $M_B = \frac{1}{N}\sum_{i=1}^{N}(M_i - O_i) = \overline{M} - \overline{O}$ |
| Mean Absolute Gross Error | $M_{AGE} = \frac{1}{N}\sum_{i=1}^{N}\lvert M_i - O_i \rvert$ |
| Root Mean Square Error | $R_{MSE} = [\frac{1}{N}\sum_{i=1}^{N}(M_i - O_i)^2]^{\frac{1}{2}}$ |
| **(3) Relative difference** | |
| Mean Normalized Bias | $M_{NB} = \frac{1}{N}\sum_{i=1}^{N}(\frac{M_i - O_i}{O_i}) \times 100\% = (\frac{1}{N}\sum\frac{M_i}{O_i} - 1) \times 100\%$ |
| Mean Normalized Gross Error | $M_{NGE} = \frac{1}{N}\sum_{i=1}^{N}(\frac{\lvert M_i - O_i \rvert}{O_i}) \times 100\%$ |
| Normalized Mean Bias | $N_{MB} = \frac{\sum_{i=1}^{N}(M_i - O_i)}{\sum_{i=1}^{N}O_i} \times 100\% = (\frac{\overline{M}}{\overline{O}} - 1) \times 100\%$ |
| Normalized Mean Error | $N_{ME} = \frac{\sum_{i=1}^{N}\lvert M_i - O_i \rvert}{\sum_{i=1}^{N}O_i} \times 100\% = \frac{M_{AGE}}{\overline{O}} \times 100\%$ |
| Fractional Bias | $F_B = \frac{1}{N}\sum_{i=1}^{N}\frac{(M_i - O_i)}{\frac{(M_i + O_i)}{2}}$ |
| Fractional Gross Error | $F_{GE} = \frac{1}{N}\sum_{i=1}^{N}\frac{\lvert M_i - O_i \rvert}{\frac{(M_i + O_i)}{2}}$ |

* $\overline{M} = \frac{1}{N}\sum_{i=1}^{N}M_i$ , $\overline{O} = \frac{1}{N}\sum_{i=1}^{N}O_i$ , $M_i$ and $O_i$ are values of model (prediction) and observation at $I$, respectively. N is number of samples (by time and/or location).

Table 2. Results of different metrics in Table 1 for different combinations of dataset in Figure 1.

| Combination* | 1 | 2 | 3 | 4 | 1+3 | 1+4 | 2+3 | 2+4 | 1+2+3+4 |
|---|---|---|---|---|---|---|---|---|---|
| $\overline{O}$ | 1.92 | 2.15 | 2.11 | 0.88 | 2.00 | 1.45 | 2.13 | 1.36 | 1.72 |
| $\overline{M}$ | 0.42 | 1.58 | 2.94 | 2.88 | 1.49 | 1.54 | 2.39 | 2.39 | 1.88 |
| $N$ | 903 | 450 | 663 | 755 | 1566 | 1658 | 1113 | 1205 | 2771 |
| $r$ | 0.79 | 0.97 | 0.97 | 0.90 | 0.54 | 0.32 | 0.90 | 0.63 | 0.51 |
| **Difference** | | | | | | | | | |
| $M_B$ | -1.50 | -0.57 | 0.83 | 1.99 | -0.52 | 0.09 | 0.26 | 1.04 | 0.16 |
| $M_{AGE}$ | 1.50 | 0.57 | 0.83 | 1.99 | 1.22 | 1.73 | 0.72 | 1.46 | 1.32 |
| $R_{MSE}$ | 4.25 | 1.07 | 1.29 | 2.70 | 3.33 | 3.62 | 1.20 | 2.23 | 2.91 |
| **Relative Difference** | | | | | | | | | |
| $M_{NB}$ | -0.82 | -0.27 | 0.43 | 4.27 | -0.29 | 1.50 | 0.14 | 2.57 | 0.96 |
| $M_{NGE}$ | 0.82 | 0.27 | 0.43 | 4.27 | 0.65 | 2.39 | 0.36 | 2.78 | 1.58 |
| $N_{MB}$ | -0.78 | -0.26 | 0.39 | 2.25 | -0.26 | 0.06 | 0.12 | 0.76 | 0.09 |
| $N_{ME}$ | 0.78 | 0.26 | 0.39 | 2.25 | 0.61 | 1.19 | 0.34 | 1.07 | 0.77 |
| $F_B$ | -1.43 | -0.33 | 0.33 | 1.12 | -0.68 | -0.27 | 0.06 | 0.58 | -0.13 |
| $F_{GE}$ | 1.43 | 0.33 | 0.33 | 1.12 | 0.96 | 1.29 | 0.33 | 0.83 | 0.90 |
| $M_{NFB}$ | -36.67 | -0.43 | 0.43 | 4.27 | -20.96 | -18.02 | 0.08 | 2.52 | -10.75 |
| $M_{NGFE}$ | 36.67 | 0.43 | 0.43 | 4.27 | 21.32 | 21.91 | 0.43 | 2.84 | 13.28 |
| $N_{MBF}$ | **-3.58** | **-0.36** | **0.39** | **2.25** | **-0.35** | **0.06** | **0.12** | **0.76** | **0.09** |
| $N_{MEF}$ | **3.58** | **0.36** | **0.39** | **2.25** | **0.82** | **1.19** | **0.34** | **1.07** | **0.77** |

* Combinations 1, 2, 3, and 4 represent the data in regions 1, 2, 3, and 4 of Figure 1, respectively. Combination "1+3" represents the data in region 1+ data in region 3 in Figure 1.

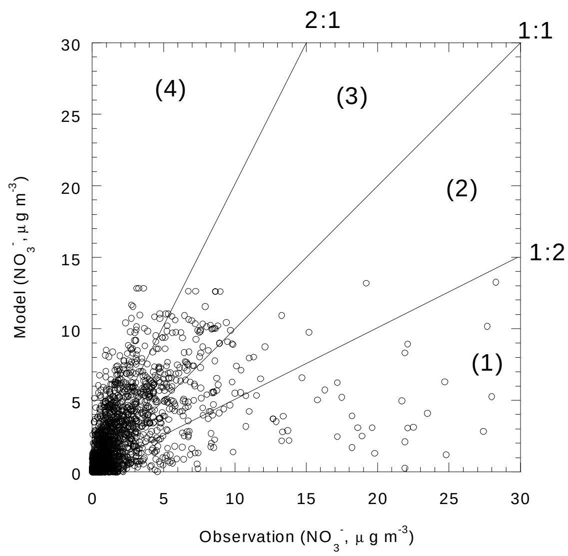


Figure 1. Comparison of modeled and observed aerosol $NO_3^-$ concentration over the continental US (see text explanation).

Table 3. Quantitative operational evaluation of CMA on the $SO_4^{2-}$ and $NO_3^-$ in 1999 summer and 2002 winter over the US for different networks

| Network | SEARCH | CASTNet | | IMPROVE | | STN |
|---|---|---|---|---|---|---|
| Year | 1999 | 1999 | 2002 | 1999 | 2002 | 2002 |
| **$SO_4^{2-}$** | | | | | | |
| Model Mean ($\overline{M}$) | 5.65 | 4.92 | 1.76 | 2.43 | 1.13 | 1.88 |
| Observation Mean ($\overline{O}$) | 4.78 | 4.57 | 1.71 | 2.07 | 0.93 | 2.23 |
| $N$ | 226 | 265 | 413 | 424 | 729 | 1149 |
| $r$ | 0.74 | 0.89 | 0.84 | 0.89 | 0.86 | 0.67 |
| $M_B$ | 0.87 | 0.36 | 0.06 | 0.36 | 0.19 | -0.34 |
| $M_{AGE}$ | 2.17 | 1.22 | 0.41 | 0.97 | 0.43 | 0.79 |
| $N_{MBF}$ | 0.18 | 0.08 | 0.03 | 0.17 | 0.22 | -0.19 |
| $N_{MEF}$ | 0.45 | 0.27 | 0.24 | 0.47 | 0.46 | 0.42 |
| **$NO_3^-$** | | | | | | |
| Mean Model ($\overline{M}$) | 0.215 | 0.32 | 2.19 | 0.16 | 0.90 | 3.38 |
| Mean OBS ($\overline{O}$) | 0.421 | 0.50 | 1.38 | 0.27 | 0.68 | 3.35 |
| $N$ | 226 | 265 | 4.15 | 424 | 689 | 1044 |
| $r$ | 0.43 | 0.28 | 0.76 | 0.31 | 0.54 | 0.36 |
| $M_B$ | -0.21 | -0.18 | 0.81 | -0.11 | 0.22 | 0.03 |
| $M_{AGE}$ | 0.33 | 0.37 | 1.11 | 0.27 | 0.67 | 2.43 |
| $N_{MBF}$ | -0.96 | -0.56 | 0.59 | -0.73 | 0.32 | 0.01 |
| $N_{MEF}$ | 1.53 | 1.16 | 0.80 | 1.70 | 0.99 | 0.72 |

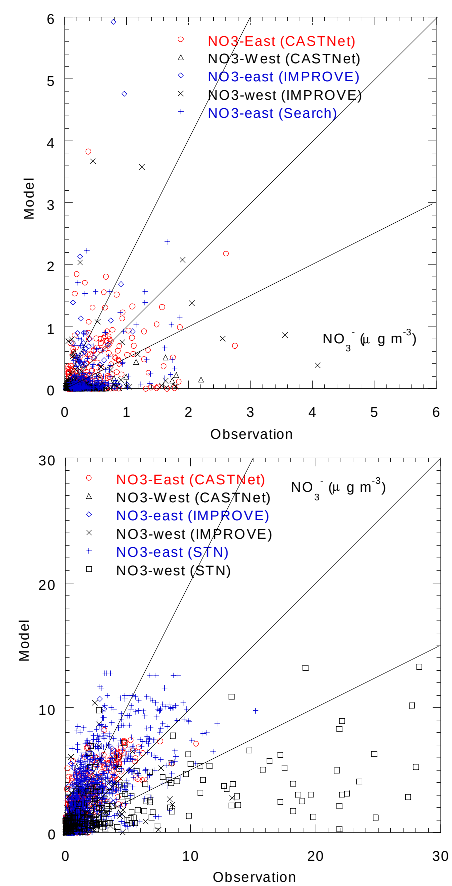


Figure 3. Scatter plots of $PM_{2.5}$ $NO_3^-$ between the model and observation over the continental US for different networks in 1999 summer (upper) and 2002 winter (lower). The 1:1, 2:1, and 1:2 lines are shown for reference.