ATMOSPHERIC SCIENCE LETTERS
Atmos. Sci. Let. **7**: 26–34 (2006)
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI:** 10.1002/asl.125

New unbiased symmetric metrics for evaluation of air quality models

Shaocai Yu,[1,†]* Brian Eder,[1‡] Robin Dennis,[1‡] Shao-Hang Chu[2] and Stephen E. Schwartz[3]

[1] *Atmospheric Sciences Modeling Division, National Exposure Research Laboratory, U.S. Environmental Protection Agency, RTP, NC 27711, USA*
[2] *Office of Air Quality Planning and Standards, U.S. EPA, RTP, NC 27711, USA*
[3] *Atmospheric Sciences Division, Brookhaven National Laboratory, Upton, NY 11973, New York*

Correspondence to: Shaocai Yu, Atmospheric Sciences Modeling Division, National Exposure Research Laboratory, U.S. Environmental Protection Agency, RTP, NC 27711, USA. E-mail: yu.shaocai@epa.gov

[†] *On assignment from Science and Technology Corporation, Hampton, VA 23666, USA.*

[‡] *On assignment from National Oceanic and Atmospheric Administration, RTP, NC 27711, USA.*

Received: 9 September 2005
Revised: 13 February 2006
Accepted: 13 February 2006

Abstract

Unbiased symmetric metrics to quantify the relative bias and error between modeled and observed concentrations, based on the factor between measured and observed concentrations, are introduced and compared to conventionally employed metrics. Application to the evaluation of several data sets shows that the new metrics overcome concerns with the conventional metrics and provide useful measures of model performance. Copyright © 2006 Royal Meteorological Society

Keywords: **unbiased symmetric metrics; evaluation; air quality model; factor**

1. Introduction

The use of models in the simulation of air quality has seen a rapid increase over the past two decades in not only the incidence of application but also the scope of that application. Once used primarily for atmospheric research, these models have had increasing utility in regulatory application and, most recently, in air quality forecasting. Regardless of the application, it is essential that these models be evaluated against measurements in order to characterize their performances so that confidence can be developed within both the air quality regulatory and air quality forecasting communities. The U.S. Environmental Protection Agency (EPA, 1991) has developed guidelines, based on Tesche *et al.* (1990), for a minimum set of statistical measures to be used for operational evaluation. Taylor (2001) proposed a graphical method to summarize multiple aspects of model performance. Operational evaluations of different air quality models in the past have yielded an array of statistical metrics that are so diverse and numerous that it is difficult to judge the overall performance of the models (Chang and Hanna, 2004; EPA, 1991; Cox and Tikvart, 1990; Seigneur *et al.*, 2000; Taylor, 2001; Yu *et al.*, 2003). Additionally, some of these metrics are inherently deficient in that they are subject to asymmetry and/or bias. In this study, a new set of unbiased symmetric metrics for the operational evaluation is proposed and applied. These new metrics, which are based on the intuitive and commonly used concept of the factor by which the modeled and observed quantities differ, provide statistical measures of that factor both as an unsigned quantity that gives its mean magnitude and as a signed quantity that gives both the mean magnitude of the factor and its sense – modeled greater or less than measured.

2. An examination of traditional evaluation metrics

A review of the literature (Chang and Hanna, 2004; EPA, 1984, 1991; Fox, 1981; Willmott, 1982; Cox and Tikvart, 1990; Weil *et al.*, 1992; Seigneur *et al.*, 2000; Yu *et al.*, 2003) reveals a plethora of metrics (summarized in Table I) used to quantify the differences between simulations and observations. Each of these metrics assumes the existence of a number N of pairs of modeled and observed concentrations M_i and O_i; the index i might be over time series at a given location, or over locations in a given spatial domain, or both. Two of the more commonly used metrics used to quantify the departure between modeled and observed quantities are the mean bias B_{MB} and the mean absolute gross error E_{MAGE} (see definitions in Table I). The mean bias is a useful measure of the overall

Table I. Summary of quantitative metrics commonly used in the operational evaluation of air quality model

Metrics	**Mathematical expression***	**Range**
(1) Correlation		
Correlation coefficient	$r = \dfrac{\sum (M_i - \overline{M})(O_i - \overline{O})}{\left\{\sum (M_i - \overline{M})^2 \sum (O_i - \overline{O})^2\right\}^{\frac{1}{2}}}$	-1 to $+1$
(2) Difference		
Mean bias	$B_{\mathrm{MB}} = \dfrac{1}{N}\sum (M_i - O_i) = \overline{M} - \overline{O}$	$-\overline{O}$ to $+\infty$
Mean absolute gross error	$E_{\mathrm{MAGE}} = \dfrac{1}{N}\sum \lvert M_i - O_i\rvert$	0 to $+\infty$
Root mean square error	$E_{\mathrm{RMSE}} = \left[\dfrac{1}{N}\sum (M_i - O_i)^2\right]^{\frac{1}{2}}$	0 to $+\infty$
(3) Relative difference		
Mean normalized bias	$B_{\mathrm{MNB}} = \dfrac{1}{N}\sum\left(\dfrac{M_i - O_i}{O_i}\right) = \left(\dfrac{1}{N}\sum\dfrac{M_i}{O_i} - 1\right)$	-1 to $+\infty$
Mean normalized absolute error	$E_{\mathrm{MNAE}} = \dfrac{1}{N}\sum\left(\dfrac{\lvert M_i - O_i\rvert}{O_i}\right)$	0 to $+\infty$
Normalized mean bias	$B_{\mathrm{NMB}} = \dfrac{\sum (M_i - O_i)}{\sum O_i} = \left(\dfrac{\overline{M}}{\overline{O}} - 1\right)$	-1 to $+\infty$
Normalized mean absolute error	$E_{\mathrm{NMAE}} = \dfrac{\sum \lvert M_i - O_i\rvert}{\sum O_i} = \dfrac{E_{\mathrm{MAGE}}}{\overline{O}}$	0 to $+\infty$
Fractional bias	$B_{\mathrm{FB}} = \dfrac{1}{N}\sum\dfrac{(M_i - O_i)}{(M_i + O_i)/2}$	-2 to $+2$
Fractional absolute error	$E_{\mathrm{FAE}} = \dfrac{1}{N}\sum\dfrac{\lvert M_i - O_i\rvert}{(M_i + O_i)/2}$	0 to 2

*$\overline{M} = \frac{1}{N}\sum M_i, \overline{O} = \frac{1}{N}\sum O_i$.

over- or underestimation by the model; the quantity is expressed in the units of the measurement (e.g. μg m^{-3}) making it useful especially for considerations of air quality. Measures other than the bias are useful to characterize the spread of the departure between the model and observations, analogous to the standard deviation of the departure in addition to the mean departure. For this reason, alternative metrics such as the mean absolute gross error E_{MAGE} are commonly employed in addition to the bias.

It is also frequently desirable to provide a measure of the relative or fractional difference between the model estimations and observations; this is generally achieved through some sort of normalization. Relative measures are particularly useful in comparing the performance of models for different substances for which concentrations are normally quite different. Historically, most such relative differences are normalized by the observed quantities. Examples include: the mean normalized bias (B_{MNB}), the mean normalized absolute error (E_{MNAE}), the normalized mean bias (B_{NMB}) and the normalized mean absolute error (E_{NMAE}) (see Table I for definitions). There are two concerns associated with these approaches to normalization that can result in misleading conclusions. This first concern is *asymmetry*. The values of both B_{MNB} and B_{NMB} can grow disproportionately as a consequence of the fact that model overestimates are unbounded whereas underestimates (for quantities such as concentrations) are bounded by – 100%. The second concern is *inflation*. The values of both B_{MNB} and E_{MNAE} can be greatly inflated by a few instances in which the observed quantity in the denominator of the expression is quite low relative to the bulk of the observations. Such a situation is not uncommon, especially when dealing with particulate matter and/or toxins. The asymmetry issue has been addressed by the introduction of the fractional bias B_{FB} and fractional absolute error E_{FAE} (Seigneur *et al.*, 2000; see Table I). Although B_{FB} and E_{FAE} can overcome the problem of asymmetry between model over- and underestimation, the significance of the metrics B_{FB} and E_{FAE} is confounded because the modeled quantity is not evaluated against the observed quantity alone, but rather against an average of observed and modeled quantities. This approach thus deviates from the traditional concept of evaluation in which the observations are considered truth. A further concern is that the scales of B_{FB} and E_{FAE} are seriously compressed beyond ±1 as B_{FB} and E_{FAE} are bounded by -2 and $+2$, and by 0 and $+2$, respectively.

These considerations have prompted the definition of new, symmetric, unbiased metrics of model performance that may be suitable for evaluations of the skill of air quality models and for the comparison of the skill of multiple models.

3. Development of new metrics

In this study, we introduce new metrics that overcome the asymmetry problem between overestimation and underestimation. These metrics are based on the intuitive and commonly used factor F_i between the observed and modeled quantity. Specifically, F_i is defined here as the ratio of modeled quantity to observed quantity if the modeled quantity exceeds the observed, whereas it is defined as the negative of the ratio of observed to modeled quantity if the observed quantity exceeds the modeled, i.e. $F_i = M_i/O_i$if $M_i \geq O_i$ and $F_i = -O_i/M_i$ if $M_i < O_i$. Note that the magnitude of F_i is always greater than or equal to unity and that the sign of F_i gives the sense of the departure: positive denotes modeled quantity greater than observed and negative denotes modeled less than observed. According to this definition $F_i = 1$ denotes perfect agreement; $F_i = 2$ denotes the model is a factor of 2 greater than observation; $F_i = -2$ denotes the model is a factor of 2 less than observation.

Following this concept, the mean normalized factor bias (B_{MNFB}), the mean normalized absolute factor error (E_{MNAFE}), the normalized mean bias factor (B_{NMBF}) and the normalized mean absolute error factor (E_{NMAEF}) are proposed and defined for a number N of pairs of modeled and observed concentrations M_i and O_i:

$$B_{\text{MNFB}} = \frac{1}{N}\sum G_i,\text{ where } G_i = \left(\frac{M_i}{O_i} - 1.0\right)$$
$$\text{if } M_i \geq O_i \text{ and } G_i = \left(1.0 - \frac{O_i}{M_i}\right) \text{ if } M_i < O_i \tag{1}$$

$$E_{\text{MNAFE}} = \frac{1}{N}\sum |G_i| \tag{2}$$

$$B_{\text{NMBF}} = \frac{\sum M_i}{\sum O_i} - 1 = \frac{\sum (M_i - O_i)}{\sum O_i}$$
$$= \frac{\overline{M}}{\overline{O}} - 1,\ \text{if } \overline{M} \geq \overline{O},\ \text{and}$$
$$= \left(1 - \frac{\sum O_i}{\sum M_i}\right) = \frac{\sum (M_i - O_i)}{\sum M_i}$$
$$= \left(1 - \frac{\overline{O}}{\overline{M}}\right),\ \text{if } \overline{M} < \overline{O} \tag{3}$$

$$E_{\text{NMAEF}} = \frac{\sum |M_i - O_i|}{\sum O_i} = \frac{E_{\text{MAGE}}}{\overline{O}}\ \text{if } \overline{M} \geq \overline{O},\ \text{and}$$
$$= \frac{\sum |M_i - O_i|}{\sum M_i} = \frac{E_{\text{MAGE}}}{\overline{M}},\ \text{if } \overline{M} < \overline{O} \tag{4}$$

where $\overline{M} = \frac{1}{N}\sum M_i$, and $\overline{O} = \frac{1}{N}\sum O_i$. In B_{MNFB} the terms that comprise the sum are positive if $M_i \geq O_i$ and negative if $M_i < O_i$. The values of B_{MNFB} and B_{NMBF} are not bounded (range from $-\infty$ to $+\infty$). The values of E_{MNAFE} and E_{NMAEF} range from 0 to $+\infty$. The above equations can be rewritten in a form that can be conveniently used to code a program when these metrics are applied making use of the quantities $S_i \equiv (M_i - O_i)/|M_i - O_i|$ and $S \equiv (\overline{M} - \overline{O})/|\overline{M} - \overline{O}|$, which denote the sense of the ratio between the modeled and observed quantities; S_i is equal to $+1$ or -1, depending on whether $M_i > O_i$ or $M_i < O_i$, respectively, and similarly for S. Thus

$$B_{\text{MNFB}} = \frac{1}{N}\sum S_i \left[\exp\left(\left|\ln\left(\frac{M_i}{O_i}\right)\right|\right) - 1\right] \tag{5}$$

$$E_{\text{MNAFE}} = \frac{1}{N}\sum |\exp(|\ln(M_i/O_i)|) - 1| \tag{6}$$

$$B_{\text{NMBF}} = S\left[\exp\left(\left|\ln\frac{\sum M_i}{\sum O_i}\right|\right) - 1\right]$$
$$= S\left[\exp\left(|\ln \overline{M}/\overline{O}\right)| - 1\right] \tag{7}$$

$$E_{\text{NMAEF}} = \frac{\sum |M_i - O_i|}{\left(\sum O_i\right)^{[1+S]/2}\left(\sum M_i\right)^{[1-S]/2}} \tag{8}$$

In Equation (8) the exponents $[1 + S]/2$ and $[1 - S]/2$ select which of the two quantities is to appear in the denominator: for $S = 1$ or -1, $[1 + S]/2 = 1$ or 0, respectively, and conversely for $[1 - S]/2$. As with the B_{MNB} and E_{MNAE}, both B_{MNFB} and E_{MNAFE} exhibit another general problem when observed values (denominator) are very small, resulting in the inflation of these metrics.

The above formulas for B_{NMBF} and E_{NMAEF} can be rewritten as follows:

For the $\overline{M} \geq \overline{O}$ case (i.e. overestimation):

$$B_{\text{NMBF}} = \frac{\sum M_i}{\sum O_i} - 1 = \frac{\sum (M_i - O_i)}{\sum O_i}$$
$$= \sum\left[\frac{O_i}{\sum O_i}\frac{(M_i - O_i)}{O_i}\right] \tag{9}$$

$$E_{\text{NMAEF}} = \frac{\sum |M_i - O_i|}{\sum O_i}$$
$$= \sum\left[\frac{O_i}{\sum O_i}\frac{|M_i - O_i|}{O_i}\right] \tag{10}$$

For the $\overline{M} < \overline{O}$ case (i.e. underestimation):

$$B_{\mathrm{NMBF}} = 1 - \frac{\sum O_i}{\sum M_i} = \frac{\sum (M_i - O_i)}{\sum M_i}$$

$$= \sum \left[\frac{M_i}{\sum M_i} \frac{(M_i - O_i)}{M_i} \right] \quad (11)$$

$$E_{\mathrm{NMAEF}} = \frac{\sum |M_i - O_i|}{\sum M_i}$$

$$= \sum \left[\frac{M_i}{\sum M_i} \frac{|M_i - O_i|}{M_i} \right] \quad (12)$$

These equations indicate that if $\overline{M} \geq \overline{O}$, B_{NMBF} and E_{NMAEF} are identical with B_{NMB} and E_{NMAE}, respectively. Equations (9) and (10) show that B_{NMBF} and E_{NMAEF} are actually the result of summing the individual mean normalized factor biases (B_{MNFB}) and errors (E_{MNAFE}) with the observed concentrations as a weighting function, respectively. For the case of $\overline{M} \leq \overline{O}$ (i.e. underestimation case), Equations (11) and (12) show that B_{NMBF} and E_{NMAEF} are the results of summing the individual mean normalized factor biases (B_{MNFB}) and errors (E_{MNAFE}) with the modeled concentrations as a weighting function, respectively. B_{NMBF} and E_{NMAEF} have the advantage of both avoiding inflation due to low values of observations in normalization (like B_{NMB} and E_{NMAE}) and maintaining adequate evaluation symmetry like B_{FB} and E_{FAE}. Both B_{NMBF} and E_{NMAEF} are also much easier to interpret than B_{FB} and E_{FAE}. For example, B_{NMBF} can be interpreted as follows: if B_{NMBF} is *positive*, the model *overestimates* the observations by a factor of $B_{\mathrm{NMBF}} + 1$; e.g. for $B_{\mathrm{NMBF}} = 1.2$, the model overestimates the observations by a factor of 2.2. If B_{NMBF} is *negative*, the model *underestimates* the observations by a factor of $1 - B_{\mathrm{NMBF}}$; for example, $B_{\mathrm{NMBF}} = -1.2$ indicates that the model underestimates the observations by a factor of 2.2. Thus, the metric B_{NMBF} indicates both the magnitude of the factor between the modeled and observed quantities and the sense of that factor (greater or less than unity). The metric E_{NMAEF} can be interpreted as follows: if $E_{\mathrm{NMAEF}} = 1.8$, this means that the absolute gross error is 1.8 times the mean observation and model prediction for overprediction ($B_{\mathrm{NMBF}} \geq 0$, or $\overline{M} \geq \overline{O}$) and underprediction ($B_{\mathrm{NMBF}} \leq 0$, or $\overline{M} \leq \overline{O}$), respectively.

4. Illustrations of the new metrics

In order to test the robustness of these new metrics against the more commonly used metrics (listed in Table I), we applied them to two different model simulations. In the first simulation, a scatter plot of the modeled *versus* observed aerosol NO_3^- concentrations was divided into four regions as shown in Figure 1 (i.e. region 1 for $0 < M_i/O_i < 0.5$, region 2 for $0.5 < M_i/O_i < 1.0$, region 3 for $1.0 < M_i/O_i \leq 2.0$ and region 4 for $2.0 < M_i/O_i$). Then, the conventionally employed metrics in Table I, along with the several new metrics, were calculated using different combinations of data in each of the four regions of Figure 1. Table II compares the several metrics of model bias and error for the several cases. For the case using only data from region 1, in which the model underestimated each of the observations by more than a factor of 2, the values of the conventional measures of model bias, the mean normalized bias B_{MNB}, the normalized mean bias B_{NMB} and the fractional bias B_{FB}, are −0.82, −0.78, −1.43, respectively. With the new metrics introduced here, the mean normalized factor bias B_{MNFB} and the normalized mean bias factor B_{NMBF} were −36.67 and −3.58, respectively. The value for B_{NMBF} (−3.58) indicates that the model underestimated the observations by a factor of 4.58 for this case, providing the most meaningful description of model performance of the several metrics. Similarly, for the case with data only in region 4, in which the model overestimated all observations by more than a factor of 2, the values of B_{MNB}, B_{NMB}, B_{FB}, B_{NMFB}, and B_{NMBF} are 4.27, 2.25, 1.06, 4.27 and 2.25, respectively. The normalized mean bias factor B_{NMBF} again provides the most meaningful description of the performance; i.e. that the model overestimated the observations by a factor of 3.25. It is especially interesting to see the results of each metric on a case combining the two regions 1 and 4, i.e. regions of substantial model underestimation and substantial overestimation. Here B_{MNB}, B_{NMB}, B_{FB}, B_{NMFB}, B_{MNFB} and B_{NMBF} are 1.50, 0.06, −0.27, 0.06, −18.02 and 0.06,

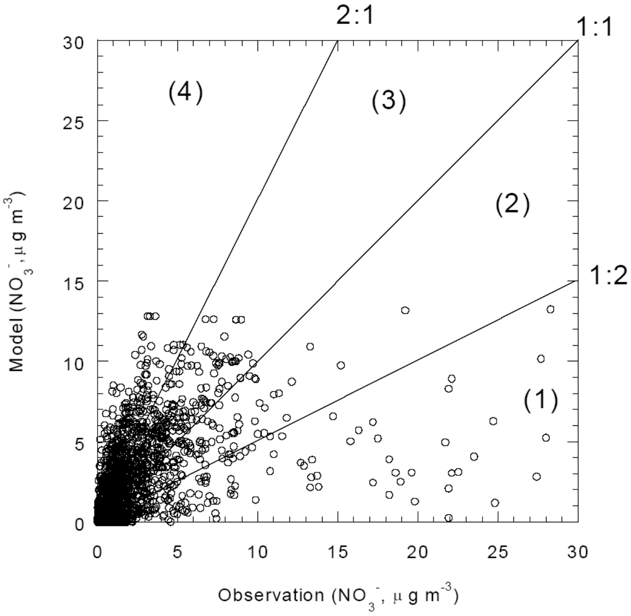

Figure 1. Comparison of modeled (M_i) and observed (O_i) aerosol NO_3^- concentrations. The 1 : 1, 2 : 1 and 1 : 2 lines are shown for reference

Table II. Results of the different metrics in Table I for different combinations of datasets in Figure 1

Combination[a]	**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									
B_{MB}	−1.50	−0.57	0.83	1.99	−0.52	0.09	0.26	1.04	0.16
E_{MAGE}	1.50	0.57	0.83	1.99	1.22	1.73	0.72	1.46	1.32
E_{RMSE}	4.25	1.07	1.29	2.70	3.33	3.62	1.20	2.23	2.91
Relative difference									
B_{MNB}	−0.82	−0.27	0.43	4.27	−0.29	1.50	0.14	2.57	0.96
E_{MNAE}	0.82	0.27	0.43	4.27	0.65	2.39	0.36	2.78	1.58
B_{NMB}	−0.78	−0.26	0.39	2.25	−0.26	0.06	0.12	0.76	0.09
E_{NMAE}	0.78	0.26	0.39	2.25	0.61	1.19	0.34	1.07	0.77
B_{FB}	−1.43	−0.33	0.33	1.12	−0.68	−0.27	0.06	0.58	−0.13
E_{FAE}	1.43	0.33	0.33	1.12	0.96	1.29	0.33	0.83	0.90
B_{MNFB}	−36.67	−0.43	0.43	4.27	−20.96	−18.02	0.08	2.52	−10.75
E_{MNAFE}	36.67	0.43	0.43	4.27	21.32	21.91	0.43	2.84	13.28
B_{NMBF}	**−3.58**	**−0.36**	**0.39**	**2.25**	**−0.35**	**0.06**	**0.12**	**0.76**	**0.09**
E_{NMAEF}	**3.58**	**0.36**	**0.39**	**2.25**	**0.82**	**1.19**	**0.34**	**1.07**	**0.77**

[a] 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 and region 3 in Figure 1.

respectively. Both B_{NMB} and B_{NMBF} show that the model slightly overestimated the observations, by a factor of 1.06, whereas the values of B_{FB} (−0.27) and B_{MNFB} (−18.02) are negative, indicating underestimation. This shows that the values of B_{FB} and B_{MNFB} can at times provide misleading (and in the case of B_{MNFB}, inflated) conclusions, in large part because of their use of both model estimations and observations in the normalization. Although the model mean (1.54 μg m^{-3}) is close to that of the observation mean (1.45 μg m^{-3}) and the values of B_{NMB} and B_{NMBF} are small (0.06), both E_{NMAE} and E_{NMAEF} (1.19) show that the absolute factor error between observations and model results is 1.19 times the mean observation. This indicates that assessment of model performance requires consideration of both relative bias (B_{NMBF}) and relative absolute error (E_{NMAEF}).

For the combination of areas 2 and 3, the values of the different metrics tend to converge; all measures of error are between 0.33 and 0.43, and all measures of bias are positive and between 0.06 and 0.14. For the entire dataset, the values of $B_{MNB}, B_{NMB}, B_{FB}, B_{MNFB}$ and B_{NMBF} are 0.96, 0.09, −0.13, −10.75 and 0.09, respectively. Both B_{NMB} and B_{NMBF} show that the mean model overestimated the mean observation by a factor of 1.09, but the values of B_{FB} and B_{MNFB} are once again negative (−0.13, −10.75) and in the case of B_{MNFB} greatly inflated.

As a second example, the metrics were applied to evaluate the performances of 11 different chemical transport models (Table III) simulating annual average concentration of nonseasalt (nss) $SO_4{}^{2-}$ at several island and coastal locations in the North and South Atlantic, as compared with measurements in Figure 2. These comparisons illustrate that conventional metrics can yield misleading results, which are overcome by the metrics introduced here. For example, the correlation coefficient r can be near unity despite systematic model underestimate (Model A); the systematic model underestimation is well captured by the metrics B_{NMBF} and E_{NMAEF}. A model such as F, which arguably does comparably to or better than Model D in capturing the observations as shown in Figure 2, exhibits much greater B_{MNB} and E_{MNAE} values as a consequence of inflation due to low observed values; in contrast, the metrics B_{NMBF} and E_{NMAEF} clearly indicate that Model F does only slightly better than Model D. For illustrative purposes, results from three fictitious model simulations were also evaluated: Model 'L' underestimates the observations by 100% (modeled concentrations are all zero); Model 'M' systematically overestimates the observations by 100% or a factor of 2; and Model 'N' assumes that all of the modeled values are $+\infty$. The conventional metrics $B_{MB}, E_{MAGE}, E_{RMSE}, B_{MNB}, E_{MNAE}, B_{NMB}$ and E_{NMAE} result in a great asymmetry between the model over- and underestimation. For example, the metric B_{NMB} is the same in magnitude, differing only in sign, for overestimation by a factor of 2 and underestimation by a factor of ∞ (model results uniformly zero) (cases M and L), despite considerable model skill in the first instance and no model skill whatsoever in the second instance. In contrast, the newly proposed statistical metrics, B_{NMBF} and E_{NMAEF}, provide much more meaningful measures of the relative performance of these models, i.e., infinite error for model estimation zero and +1 (100%) for model estimation a factor of two high. For the criteria of model performance taken as: $|B_{NMBF}| \leq 25\%$ and $E_{NMAEF} \leq 35\%$, only Models E, G, and H satisfy these criteria, with the best performance being exhibited by Model H and the worst

Table III. Results of different metrics in Table 1 for the performances of different models on non-seasalt sulfate in Figure 2

Models*	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**
$\overline{O}$	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
$\overline{M}$	0.35	1.37	1.19	1.34	1.22	1.16	1.19	1.02	0.79	1.23	0.67	0.00	1.95	$+\infty$
N	9	9	9	9	9	9	9	9	9	9	9	9	9	9
r	0.96	0.84	0.74	0.78	0.84	0.77	0.95	0.98	0.61	0.69	0.77	0.00	1.00	0.00
Difference														
B_{MB}	−0.63	0.40	0.21	0.37	0.24	0.18	0.21	0.05	−0.19	0.25	−0.31	−0.98	+0.98	$+\infty$
E_{MAGE}	0.63	0.46	0.42	0.52	0.34	0.42	0.24	0.14	0.42	0.52	0.41	0.98	+0.98	$+\infty$
E_{RMSE}	0.79	0.55	0.52	0.70	0.49	0.48	0.37	0.16	0.58	0.63	0.55	0.98	+0.98	$+\infty$
Relative Difference														
B_{MNB}	−0.65	1.23	0.91	0.38	0.70	1.40	0.34	0.33	0.19	0.75	−0.06	−1.00	+1.00	$+\infty$
E_{MNAE}	0.65	1.26	1.01	0.60	0.80	1.58	0.39	0.39	0.59	0.94	0.52	1.00	+1.00	$+\infty$
B_{NMB}	−0.64	0.41	0.22	0.38	0.25	0.18	0.21	0.05	−0.20	0.26	−0.32	−1.00	+1.00	$+\infty$
E_{NMAE}	0.64	0.47	0.43	0.53	0.34	0.43	0.25	0.15	0.44	0.53	0.42	1.00	+1.00	$+\infty$
B_{FB}	−1.00	0.53	0.37	0.16	0.30	0.35	0.22	0.16	−0.04	0.30	−0.24	−2.00	+0.67	$+\infty$
E_{FAE}	1.00	0.56	0.48	0.45	0.43	0.56	0.27	0.24	0.47	0.53	0.53	2.00	+0.67	$+\infty$
B_{MNFB}	−2.81	1.23	0.89	0.27	0.66	1.35	0.34	0.32	0.02	0.70	−0.34	$-\infty$	+1.000	$+\infty$
E_{MNAFE}	2.81	1.26	1.02	0.70	0.84	1.63	0.39	0.40	0.76	1.00	0.80	$+\infty$	+1.000	$+\infty$
B_{NMBF}	**−1.81**	**0.41**	**0.22**	**0.38**	**0.25**	**0.18**	**0.21**	**0.05**	**−0.24**	**0.26**	**−0.46**	$-\infty$	+1.000	$+\infty$
E_{NMAEF}	**1.81**	**0.47**	**0.43**	**0.53**	**0.34**	**0.43**	**0.25**	**0.14**	**0.54**	**0.53**	**0.61**	$+\infty$	+1.000	$+\infty$

* The units of $\overline{O}$, $\overline{M}$, B_{MB}, E_{MAGE} and E_{RMSE} are μg m^{-3}.

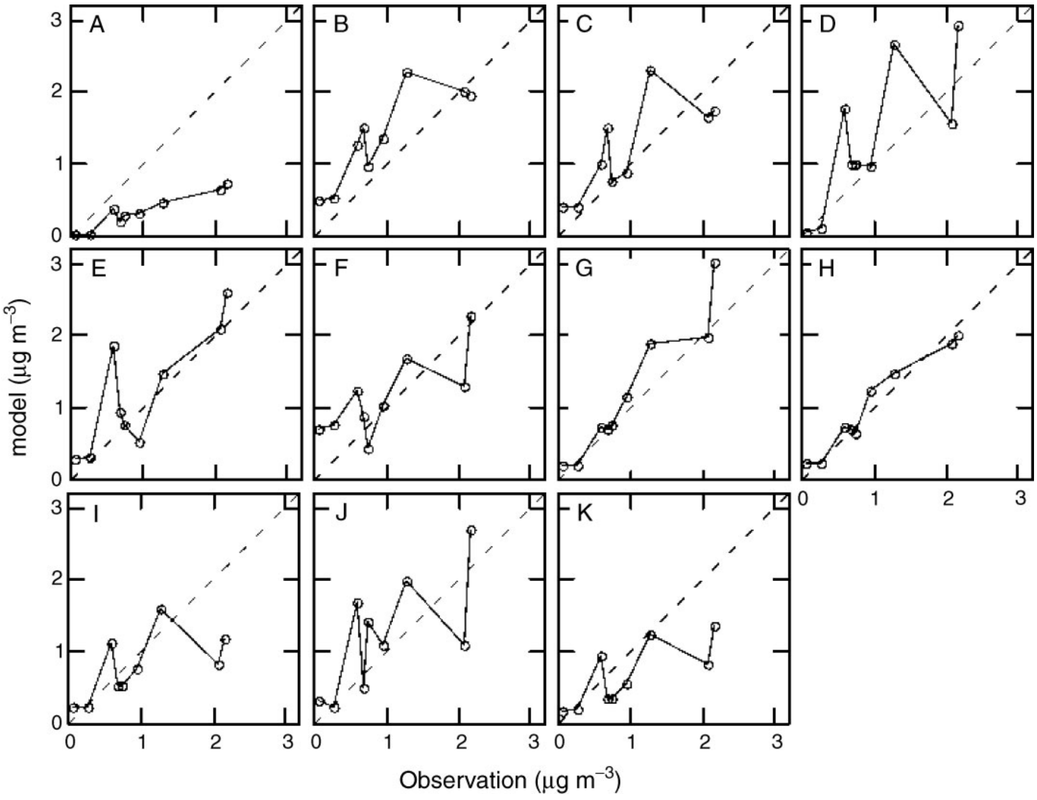

Figure 2. Comparisons of annual average concentrations of nonseasalt sulfate from 11 chemical transport models with observations at a series of island and coastal stations in the North and South Atlantic. Data are from Penner *et al.* (2001)

performance being exhibited by Model A; these metrics are consistent with the scatter plots of Figure 2.

5. Applications of new metrics using CMAQ simulations

Further illustration of the utility of the newly proposed metrics is provided for a simulation of annual mean concentrations of SO_4^{2-} and NO_3^- carried out with the U.S. EPA Models-3/Community Multiscale Air Quality (CMAQ) model (2004 release; version 4.4). Further information about the simulations, including details on the networks used in the evaluation (Clean Air Status and Trends Network (CASTNet), Interagency Monitoring of Protected Visual Environments (IMPROVE) and Speciated Trends Network

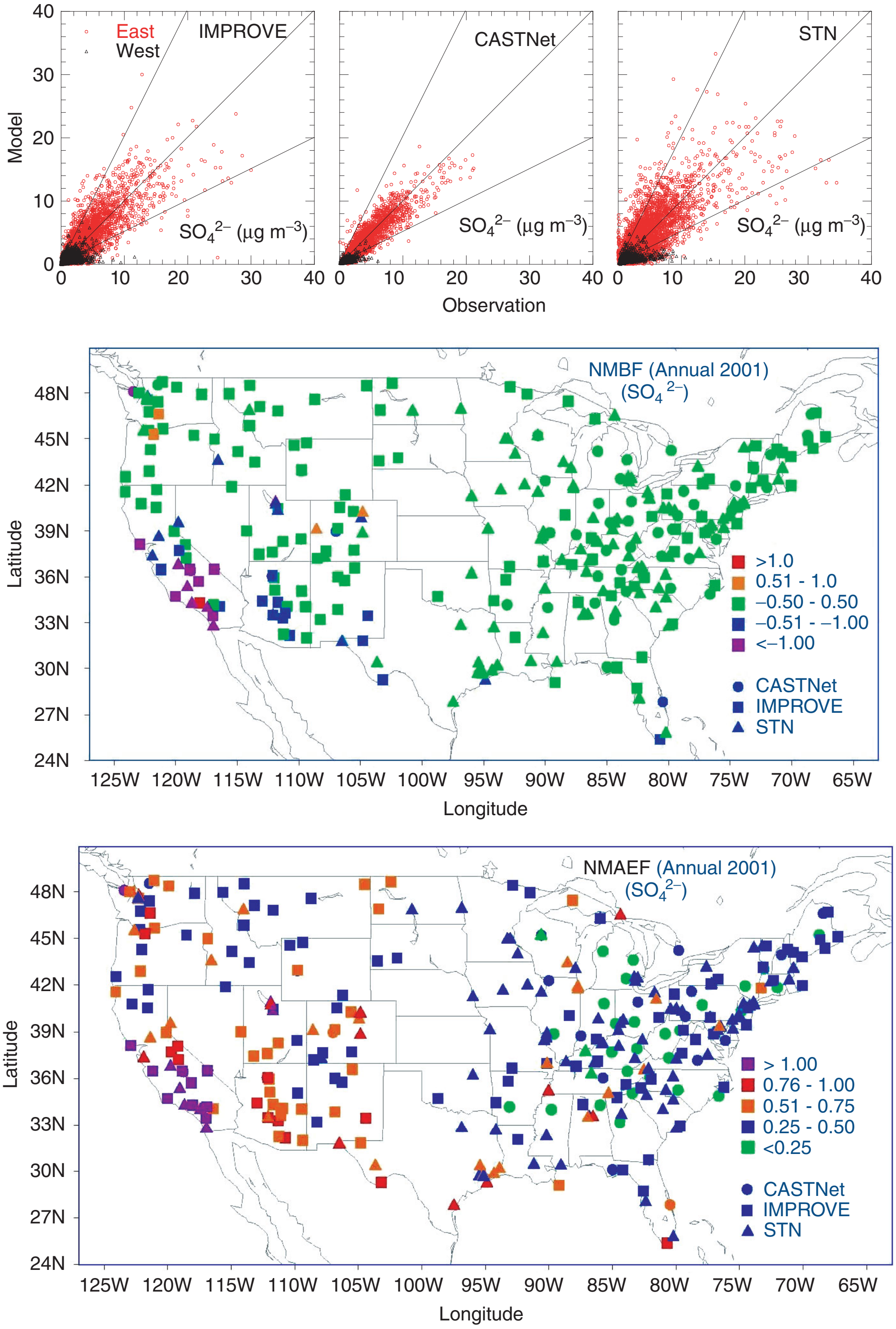

Figure 3. Scatter plot of SO_4^{2-} between the CMAQ model (M_i) and observation (O_i) (upper panel), and spatial distributions of B_{NMBF} and B_{NMAEF} over the US for different networks for 2001 simulation. The 1 : 1, 2 : 1 and 1 : 2 lines are shown for reference in the scatter plots

(STN)) can be found in Eder and Yu (2006). Table IV reveals that for SO_4^{2-} concentrations the vast majority of the simulations agree with the observations within a factor of 2 (Figure 4). The B_{NMBF} values for each of the three networks tend to be small and negative, ranging from −0.02 (STN) to −0.06 (IMPROVE) and −0.11 (CASTNet). This indicates that the CMAQ model underestimated SO_4^{2-} concentrations by factors ranging from 1.02 to 1.11. Examination of the B_{NMBF} as a function of location (Figure 3) reveals better performance over the eastern half of the domain, where the majority of the B_{NMBF} values lie within ±0.50. Performance degrades somewhat in the West, especially in California, where values of B_{NMBF} are often below −1.00, indicating that the model underestimates by more than a factor of 2.

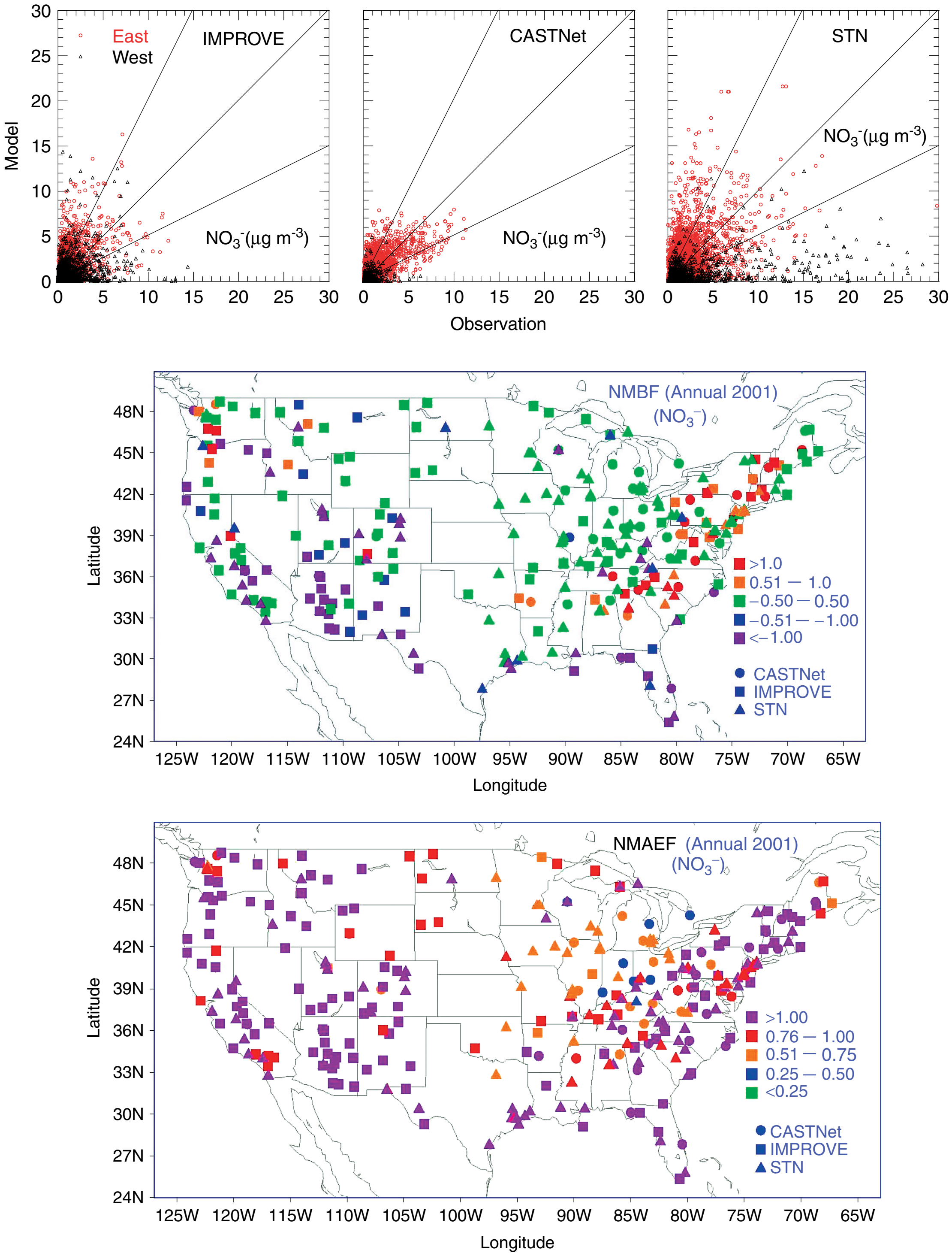

Figure 4. Same as Figure 3 but for NO_3^-

For aerosol NO_3^-, the B_{NMBF} values associated with the CASTNet and IMPROVE networks are small and positive, ranging from 0.04 (IMPROVE) to 0.05 (CASTNet). They are negative and somewhat larger for STN sites (−0.19). This indicates that CMAQ slightly overestimates NO_3^- concentrations by factors of 1.04 and 1.05 for IMPROVE and CASTNet, respectively, while underestimating against STN sites by a factor of 1.19. When examined over the spatial domain (Figure 4), large differences in performance become evident. For example, CMAQ tends to overestimate NO_3^- concentrations in the eastern portion of the domain, where B_{NMBF} often exceeds +0.50, while it tends to underestimate in most western locations, where B_{NMBF} falls below −0.50 (factors of 1.5 over- and underestimations, respectively). Exceptions to this general east *versus* west difference do exist, most notably for locations along the Gulf of Mexico, where the model underestimates by more than a factor of 2, and in Washington and Oregon, where the model overestimates. The very large values of E_{NMAEF} for aerosol

Table IV. Statistical metrics associated with an annual simulation (2001) of the 2004 release of models-3 CMAQ

	SO_4^{2-}			NO_3^-		
Network	**CASTNet**	**IMPROVE**	**STN**	**CASTNet**	**IMPROVE**	**STN**
$\overline{O}$	2.88	1.60	3.33	1.04	0.50	1.48
$\overline{M}$	3.21	1.69	3.40	0.99	0.48	1.77
N	3736	13447	6970	3735	13398	6130
r	0.92	0.85	0.77	0.67	0.52	0.37
B_{MB}	−0.32	−0.09	−0.07	0.05	0.02	−0.29
E_{MAGE}	0.80	0.66	1.43	0.70	0.46	1.42
B_{NMBF}	**−0.11**	**−0.06**	**−0.02**	**0.05**	**0.04**	**−0.19**
E_{NMAEF}	**0.28**	**0.41**	**0.43**	**0.71**	**0.94**	**0.96**

NO_3^- in Figure 4 and Table IV indicate the spread of departure between the model and observations.

6. Summary

In addition to some commonly used metrics, four new symmetric metrics are introduced, two of which (i.e. B_{NMBF} and E_{NMAEF}) are found to be statistically robust measures of the factor by which the model results differ from the observations and of the sense of that factor. These two new metrics provide readily interpretable measures of model performance, which are symmetric and avoid inflation that may be caused by low values of the observed quantities. These metrics use only observed data as the model evaluation, and thus serve as the basis for a rigorous evaluation of model performance.

7. Disclaimer

This work has been subjected to U.S. Environmental Protection Agency peer review and approved for publication. The research presented here was performed under the Memorandum of Understanding between the U.S. Environmental Protection Agency (EPA) and the U.S. Department of Commerce's National Oceanic and Atmospheric Administration (NOAA) and under agreement number DW13921548. This work constitutes a contribution to the NOAA Air Quality Program. Although it has been reviewed by EPA and NOAA and approved for publication, it does not necessarily reflect their policies or views. Work by S. E. Schwartz was supported by the US Department of Energy under Contract No DE-ACO2-98CH10886 to Brookhaven Science Associates, LLC.

References

Chang JC, Hanna S. 2004. Air quality model performance. *Meteorol. Atmos Phys.* **87**: 167–196.

Cox WM, Tikvart JA. 1990. A statistical procedure for determining the best performing air quality simulation model. *Atmospheric Environment* **24**: 2387–2395.

Eder B, Yu SC. 2006. A performance evaluation of the 2004 release of Models-3 CMAQ. *Atmospheric Environment* (in press).

EPA. 1984. Interim procedures for evaluating air quality models (revised). EPA-450/4-84-023, U.S. Environmental Protection Agency.

EPA. 1991. Guideline for regulatory application of the urban airshed model. US EPA Report No. EPA-450/4-91-013. United States EPA, Office of Air Quality Planning and Standards: Research Triangle Park, NC.

Fox DG. 1981. Judging air quality model performance. *Bulletin of the American Meteorological Society* **62**: 599–609.

Penner JE, Andreae M, Annegarn H, Barrie L, Feichter J, Hegg D, Jayaraman A, Leaitch R, Murphy D, Nganga J, Pitari G. 2001. Aerosols, their direct and indirect effects. In *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden P, Dai X, Maskell K (eds). Cambridge University Press: Cambridge; 289–348.

Seigneur C, Pun B, Pai P, Louis J-F, Solomon P, Emery C, Morris R, Zahniser M, Eorsnop D, Koutrakis P, White W, Tombach I. 2000. Guidance for the performance evaluation of three-dimensional air quality modeling systems for particulate matter and visibility. *Journal of the Air & Waste Management Association* **50**: 588–599.

Taylor KE. 2001. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research* **106**(D7): 7183–7192.

Tesche TW, Georgopolous P, Lurmann FL, Roth PM. 1990. *Improvement of Procedures for Evaluating Photochemical Models*; Report PB 91-160374; National Technical Information Service: Springfield, VA, 1990.

Weil JC, Sykes RI, Venkatram A. 1992. Evaluating air-quality models: review and outlook. *Journal of Applied Meteorology* **31**: 1121–1145.

Willmott CJ. 1982. Some comments on the evaluation of model performance. *Bulletin American Meteorological Society* **63**: 1309–1313.

Yu SC, Kasibhatla PS, Wright DL, Schwartz SE, McGraw R, Deng A. 2003. Moment-based simulation of Microphysical properties of Sulfate Aerosols in the Eastern United States: Model description, evaluation and regional analysis. *Journal of Geophysical Research* **108**(D12): 4353, DOI:10.1029/2002JD002890.