6A.3A IMPACT OF LINEAR VARIANCE CALIBRATION ON ENSEMBLE WIND VARIANCE IN ATMOSPHERIC TRANSPORT AND DISPSERSION FORECASTS

Walter C. Kolczynski, Jr.[1]*
David R. Stauffer[1]
R. Ian Sykes[2]
Robert Long[2]
Sue Ellen Haupt[1†]
Aijun Deng[1]
[1] Pennsylvania State University, University Park, PA
[2] Sage Management, Princeton, NJ
[†] Current affiliation: National Center for Atmospheric Research

1. INTRODUCTION

An accurate estimate of uncertainty is an important component in any risk assessment. With the correct uncertainty (or, more precisely, the event likelihood that can be calculated from it), decisions can better be made to minimize expected loss. In some applications, such as in atmospheric transport and dispersion (AT&D) forecasts of a biological or chemical release, this can mean minimizing expected casualties.

One of the biggest contributors to AT&D forecast uncertainty is the uncertainty in the meteorological (MET) forcing such as the wind field (Rao 2005). Since AT&D models often use the output from numerical weather prediction (NWP) models to provide the MET information, accurate estimates of the uncertainty in those mean wind forecasts contribute to the uncertainty in the resulting AT&D forecast. The Second-order Closure Integrated Puff (SCIPUFF, Sykes et al. 2006) model incorporates uncertainty in the input mean winds via the variance of each wind component (zonal, *u*, and meridional, *v*) and the covariance of the two components.

A common method for estimating the uncertainty in NWP forecasts is the use of an ensemble of forecasts, with multiple forecasts produced using a variety of initial / lateral boundary conditions and model physics / parameterizations. The goal of using MET ensembles is to span the possible outcomes given the uncertainties in the initial state and modeling system (Leith 1974).

While ensemble forecasting is a step toward estimating the uncertainty of NWP forecasts, the size of operational MET ensembles is insufficient to fully represent the probability density function (PDF) of possible forecasts. An ensemble capable of doing so is impractical with current computing resources. Therefore, any MET ensemble provides a sampling of the full forecast PDF and any measures of the uncertainty from the ensemble (such as variance) should be evaluated for applicability and calibrated if necessary.

Kolczynski et al (2009) introduced the Linear Variance Calibration (LVC) as one potential method of calibration and demonstrated its impact on one AT&D forecast case study. While that case study demonstrated both a qualitative and a quantitative improvement in the mean concentration forecast, a large sample is needed in order to properly evaluate the effectiveness of calibration on the reliability of the AT&D forecasts.

This study uses a sample of 112 24-h forecasts spread over an eight-month period from a joint NWP-SCIPUFF testbed to evaluate the impact of LVC on forecasts by the AT&D model SCIPUFF. An independent SCIPUFF simulation based upon a high-resolution NWP dynamic analysis is used as verification. Section 2 summarizes the Linear Variance Calibration and evaluation methods used in this study. Section 3 provides the details of the MET/SCIPUFF testbed and the sampling methodology used. The performance of SCIPUFF AT&D forecasts using both uncalibrated and calibrated wind variances are then evaluated in section 4. Conclusions and future work are presented in section 5.

2. METHODOLOGY

a. Linear Variance Calibration

The LVC used here is similar to that introduced in Kolczynski et al. (2009). The LVC is determined by computing a linear regression between the variance of the ensemble and the variance in the errors of the MET forecasts. Since there is only one error available for any particular forecast, forecasts with similar ensemble variances are grouped together, and the error variance is computed on the sample with similar ensemble variances.

The sample variance, $\hat{\sigma}_{ei}^2$, of the i^{th} forecast point is calculated from the ensemble members as usual:

$$\hat{\sigma}_{ei}^2 = \frac{1}{M-1}\sum_{m=1}^{M}(s_{mi} - \overline{s_i})^2 \quad (1)$$

where M is the ensemble size, s_{mi} is the forecast value for the m^{th} ensemble member, and the overbar denotes the sample mean. The error of the ensemble mean, ε_i, compared to the observed verification value, s_{vi}, is also calculated:

$$\varepsilon_i = \overline{s_i} - s_{vi} \quad (2)$$

Equations (1) and (2) provide a series of ensemble variance-error pairs, one for each forecast i within a region during some specified training period preceding the forecast. These pairs are then ordered based on

* *Corresponding Author Address:* Walter Kolczynski, Jr. Dept. of Meteorology, Pennsylvania State University, 503 Walker Bldg. State College, PA 16802; wck122@psu.edu.

their ensemble variance and grouped into equally populated bins containing N pairs each. The ensemble variances within each bin are averaged to get a representative ensemble variance of the bin $\overline{{}^{b}\hat{\sigma}_e^2}$:

$$\overline{{}^{b}\hat{\sigma}_e^2} = \frac{1}{N}\sum_{i\in b}\hat{\sigma}_{ei}^2 \quad (3)$$

Most importantly, the grouping allows us to compute an estimated actual error variance, $\overline{{}^{b}\hat{\sigma}_a^2}$, for the bin:

$$\overline{{}^{b}\hat{\sigma}_a^2} = \frac{1}{N-1}\sum_{i\in b}(\varepsilon_i - {}^{b}\bar{\varepsilon})^2 \quad (4)$$

where ${}^{b}\bar{\varepsilon}$ is the average error in bin b. With these variance pairs, we use linear regression of $\overline{{}^{b}\hat{\sigma}_a^2}$ depending on $\overline{{}^{b}\hat{\sigma}_e^2}$ to produce estimates of the slope, $\hat{\beta}_1$, and the y-intercept, $\hat{\beta}_0$.

b. Evaluation Tools

There are two primary tools that are used in our evaluation of the SCIPUFF forecasts: the reliability diagram and the cumulative rank probability score (CRPS). Our calibration technique focuses on improving the reliability of the MET winds that drive SCIPUFF, and thus it is logical to assess the reliability of the resulting SCIPUFF forecasts. CRPS is chosen as an evaluation method that combines all aspects (i.e.., reliability, resolution and observation uncertainty) of probabilistic forecast verification into a single measure.

A reliability diagram is constructed by plotting the predicted probability of a certain event occurring on the abscissa against the rate at which the event actually occurred on the ordinate. Forecasts with a similar probability of the event are grouped together to provide the sample for the verification rate. For example, with 5% bins, all forecasts having between a 0% and 5% chance of occurrence would be averaged and compared against the actual occurrence rate for those forecasts.

A perfectly reliable forecast will have the event occur as often as it is predicted to occur (e.g., the event will occur 50% of the time that it is predicted to occur 50% of the time). Forecasts plotted above the perfect reliability (y=x) line are *under-confident*, as the event occurs more often than is forecasted, and forecasts below the perfect reliability line are *over-confident*, as the event occurs less frequently than predicted. A forecast with a horizontal line is said to have no *resolution*, as events forecasted to occur at different frequencies occur at the same rate regardless of the forecast. A more complete treatment of reliability diagrams is available in Wilks (2006).

To summarize the reliability into a single metric that can be easily compared between different experiments, we also compute the mean absolute reliability error (MARE) for the reliability curve. The MARE is simply the absolute difference between the observed frequency and the predicted frequency, averaged across each probability bin with at least one forecast. Lower scores are better, and zero indicates perfect reliability.

To assess the overall performance of the SCIPUFF forecasts, we use the CRPS. The CRPS is a strictly proper score (Gneiting and Raftery 2007) that evaluates a probabilistic forecast based on the degree to which the forecast cumulative distribution function (CDF) deviates from the CDF of the verification. A scoring rule is strictly proper if it is uniquely optimized by making a forecast with the same distribution as that from which the verification is drawn. The CDF of the verification is generally a step-function at the observed value, but this is not the case for this study due to the nature of our verification. Because the CDF is used in the score, all aspects of the probabilistic forecast (resolution, reliability and observation uncertainty) contribute to the CRPS. The exact calculation of the CRPS is given by:

$$CRPS = \int_{-\infty}^{\infty}\left[P_f(y) - P_o(y)\right]^2 dy \quad (5)$$

where $P_f(y)$ is the CDF of the forecast at value y and $P_o(y)$ is the CDF of the verification at value y. The CRPS of all forecasts is averaged to provide an overall score for the forecast region and period. Values are always positive, with lower values indicating a better forecast.

3. EXPERIMENTAL DESIGN

a. MET-SCIPUFF Testbed

The testbed period extends from 1 November 2009 through 7 July 2010. The NWP forecasts used for the NWP-SCIPUFF testbed are obtained from the National Centers for Environmental Protection Short-Range Ensemble Forecast (SREF) system. During this period, all 21 SREF members are available on a 32-km grid (NCEP grid 221) that covers North America. Forecasts are available every three hours and we use the SREF initialized at 0900 UTC.

The LVC uses the North American Regional Reanalysis (NARR) as the verification. These analyses are available on the same spatial grid as the SREF. Calibration is calculated separately for the two wind components (*u* and *v*) based on the 10-m above ground level forecasts, and verification is performed for the 14 days preceding the forecast at each grid location. The covariance of *u* and *v* is adjusted to maintain a constant correlation through the calibration process. A new calibration is calculated for each three-hour forecast period.

SCIPUFF forecasts begin at 0300 UTC using the ensemble mean of the SREF forecasts initialized at 0900 UTC the previous day. Two different wind variance experiments are included. The first experiment, Uncal, uses the uncalibrated wind variance from the ensemble. The second experiment, LVC_09z, uses wind variances calibrated using LVC.

The MET information driving the SCIPUFF verification is provided by a 4-km dynamic analysis using the PSU/NCAR MM5 model and four-dimensional data assimilation (FDDA). Because MET data are assimilated throughout the simulation, this method produces a very realistic estimate of the actual state of the atmosphere for air-chemistry / AT&D model applications (e.g., Tanikulu et al. 2000, Deng et al. 2004). The dynamic analysis is allowed 12 h to spin-up

and assimilate observations before being used for the SCIPUFF verification.

Both the SCIPUFF forecasts and the SCIPUFF verification simulate the instantaneous release of a passive tracer from State College, PA beginning at 0300 UTC (2200 LST). For the SCIPUFF verification simulation, the tracer is allowed to disperse for 24 h, and SCIPUFF output is available hourly. This means that SREF forecasts of 18 - 42 h are used to drive the SCIPUFF predictions. SCIPUFF forecasts at 4, 6, 12, 18 and 24 h after release are evaluated.

b. Sampler Grid

For most of our evaluations we use a uniformly spaced grid of sampler locations. All samplers are located at ground level with a spacing of 0.05°×0.05° within the region bounded by 80.6° W, 69.7°W, 36.5°N and 45.1°N. SCIPUFF forecasts, providing both a mean value and a variance based on a clipped normal distribution, allow for full probabilistic predictions at each location. Both concentration and surface dosage (integrated concentration) predictions are available for each sampler point.

c. Adaptive Grid

In addition to the sampler grid, for dosage reliability we also consider results on an adaptive grid (Sykes 2006). The adaptive grid location and spacing of the points depend on the structure of the dosage field. Regions where surface dosage has high spatial variability have higher resolution than areas with little spatial variability. Results using the adaptive grid are weighted by the area that each sample represents. In the reliability diagram using the adaptive grid, the secondary axis indicates the total area represented by forecasts within the bin instead of the number of unique forecasts.

4. RESULTS

Figure 1 shows the reliability diagram for 4-h concentration thresholds of 10^{-14} kg·m^{-3} (blue), 10^{-12} kg·m^{-3} (green), and 10^{-10} kg·m^{-3} (red) on the uniform sampler grid. The secondary axis indicates the number of unique forecasts that contributed to each probability bin. Forecasts using uncalibrated wind variances (dotted lines, with open bars indicating unique forecasts) for the lower two concentration thresholds are under-confident below 5%, and then they are over-confident in their higher probability forecasts, with little resolution or even a negative correlation between forecast probability and observed probability. However, when LVC is applied to the wind variance used by SCIPUFF (solid lines, filled bars), these forecasts are more reliable (the line is located closer to the black-dashed perfect reliability line). At the highest threshold shown, both uncalibrated and calibrated forecasts are generally over-confident at all probability levels.

Note that the thresholds are spaced two orders of magnitude apart, so that the highest threshold is 10000 times larger than the smallest. Even with this two orders of magnitude separation, the reliability of the two lowest thresholds is very similar, both for the uncalibrated and the calibrated concentration forecasts. Experimentation shows that this is a common attribute: the reliability is only sensitive to the threshold near the upper limit of concentrations. Thresholds in between the two lowest thresholds shown, as well as those below, all give similar reliability results. Thresholds a bit higher than the highest threshold shown have few or no forecasts with probabilities above 1%, making a reliability analysis meaningless.

Finally, it can be seen that the probability where the forecasts lose resolution for all experiments is about the same probability where the number of unique forecasts contributing to the bin falls below 20.

At 12 h (Fig. 2), results are similar to those at 4 h. The lower two concentration thresholds shown at this time, 10^{-16} kg·m^{-3} (blue) and 10^{-14} kg·m^{-3} (green), exhibit similar reliabilities for both the uncalibrated and calibrated experiments, as do the lower concentrations at 4 h. Uncalibrated forecasts are under-confident for all three thresholds at lower predicted probabilities. At higher predicted probabilities, the uncalibrated forecasts become over-confident and lose resolution about where the number of unique forecasts falls below 20, just as with the 4-h reliabilities. Applying calibration to the wind variances results in generally more reliable forecasts, with forecasts that are less under-confident at lower probabilities and forecasts that are less over-confident at upper probabilities.

The MARE for five different forecast lead times (4, 6, 12, 18 and 24 h) and three different thresholds are shown in Fig. 3. The MARE for the calibrated forecasts, shown in teal, are better than those for the uncalibrated forecasts (black) at every lead time / concentration threshold studied except one. This demonstrates that calibration is improving the reliability of concentration forecasts throughout the SCIPUFF forecast. Since our qualitative evaluation indicates sensitivity to the number of unique forecasts included in each bin, we also calculated the MARE using only bins to which at least 20 different forecasts contributed. These results are shown in Fig. 4. Even limiting MARE to include only bins with 20 different forecasts, the calibrated forecasts have lower MAREs than the uncalibrated forecasts.

Calibration also provides improvement in the CRPS of concentration forecasts at all five of the forecast lead times studied (Fig. 5). These improved CRPSs are evidence that the improvement in the reliability is improving the overall probabilistic concentration forecast.

The reliability of the surface dosage forecasts also improves when calibration is applied to the wind variances. Figure 6 shows the reliability for three dosage thresholds of 10^{-11} kg·s·m^{-3} (blue), 10^{-9} kg·s·m^{-3} (green), and 10^{-7} kg·m^{-3} (red) for 4-h forecasts. As seen with concentration forecast reliability, dosage reliabilities for thresholds below the highest relevant thresholds are similar, as evidenced by the similar behavior in the lower two thresholds in Fig. 6. At the lower predicted probabilities, uncalibrated forecasts of all three dosage

thresholds shown are under-confident. At higher predicted probabilities, uncalibrated forecasts are over-confident for all three dosage thresholds. Calibration improves the reliability in both of these regions, resulting in forecasts that are less under-confident at lower probabilities and forecasts that are less over-confident at higher probabilities.

The MARE statistics presented in Fig. 7 show that this improvement in dosage reliability on the regular sampler grid appears to be reduced and reversed as forecast lead time increases. The MARE deterioration occurs for calibrated forecasts at higher dosage thresholds first, and then it spreads to lower thresholds as lead time increases further. By comparison, the surface dosage reliability at 24 h calculated using the adaptive grid at higher dosages (Fig. 8) shows qualitative improvement. It is currently unclear why calibration should degrade some of the sampler grid surface dosage results when the sampler grid concentration forecasts are almost universally improved in the same statistics when using calibration. It is reassuring that the adaptive grid results in Fig. 8 show qualitative improvement when using calibration for high dosages at 24 h. Further investigation into the sampler grid dosage results is needed.

5. CONCLUSIONS

This study evaluates the impact of applying Linear Variance Calibration (LVC) to the MET input wind variances on resulting probabilistic atmospheric transport and dispersion (AT&D) forecasts using the SCIPUFF model. This goal is accomplished by using a joint MET / SCIPUFF testbed where SCIPUFF forecasts using MET ensemble forecasts as input are compared to SCIPUFF simulations driven by a high-resolution MET dynamic analysis throughout the period.

It is demonstrated that LVC improves probabilistic forecasts of concentration on a fixed sampler grid at 4 h, 6h, 12 h and 24 h compared to similar forecasts using uncalibrated wind variances, as measured by both the reliability at several concentration thresholds and the cumulative rank probability score (CRPS).

Surface dosage results on the fixed sampler grid show mixed results for the effect of LVC while surface dosage reliability calculated on an adaptive grid that adjusts spatial resolution on the dosage field shows qualitative improvement at 24 h.

Acknowledgements. This work was supported by the Defense Threat Reduction Agency (DTRA) under the supervision of Dr. John Hannan via Global Security & Engineering Solutions, a Division of L-3 Services, Inc., subcontract 2008-1113, Prime Contract # DTRA01-03-D-0013.

6. REFERENCES

Deng, A., N.L. Seaman, G.K. Hunter, and D.R. Stauffer, 2004: Evaluation of interregional transport using the MM5-SCIPUFF system. *Journal of Applied Meteorology*, **43,** 1864–1886.

Gneiting, T. and A.E. Raftery, 2007. Strictrly Proper Scoring Rules, Prediction and Estimation. *Journal of the American Statistical Association*, **102**, 359-378.

Kolczynski, Walter C. Jr., D.R. Stauffer, S.E. Haupt and A. Deng, 2009: Ensemble Variance Calibration for Representing Meteorological Uncertainty for Atmospheric Transport and Dispersion Modeling. *Journal of Applied Meteorology and Climatology*, **48**, 2001-2021.

Leith, C.E., 1974: Theoretical Skill of Monte Carlo Forecasts. *Monthly Weather Review*, **102**, 409-418.

Rao, K.S., 2005: Uncertainty Analysis in Atmospheric Dispersion Modeling. *Pure and Applied Geophysics*, **162**, 1893-1917.

Sykes, R.I., S.F. Parker, D.S. Henn and B. Chowdhury 2006: SCIPUFF Version 2.2, Technical Documentation, A.R.A.P. Report no. 729, L-3 Titan Corp., Princeton, NJ, 317 pp.

Tanrikulu, S., D.R. Stauffer, N.L. Seaman, and A.J. Ranzieri, 2000. A Field-Coherence Technique for Meteorological Field-Program Design for Air Quality Studies. Part II: Evaluation in the San Joaquin Valley. *Journal of Applied Meteorology*, **39**, 317-334.

Wilks, D.S., 2006: *Statistical Methods in the Atmospheric Sciences, Second Ed.* Academic Press, Burlington, MA, 627 pp.

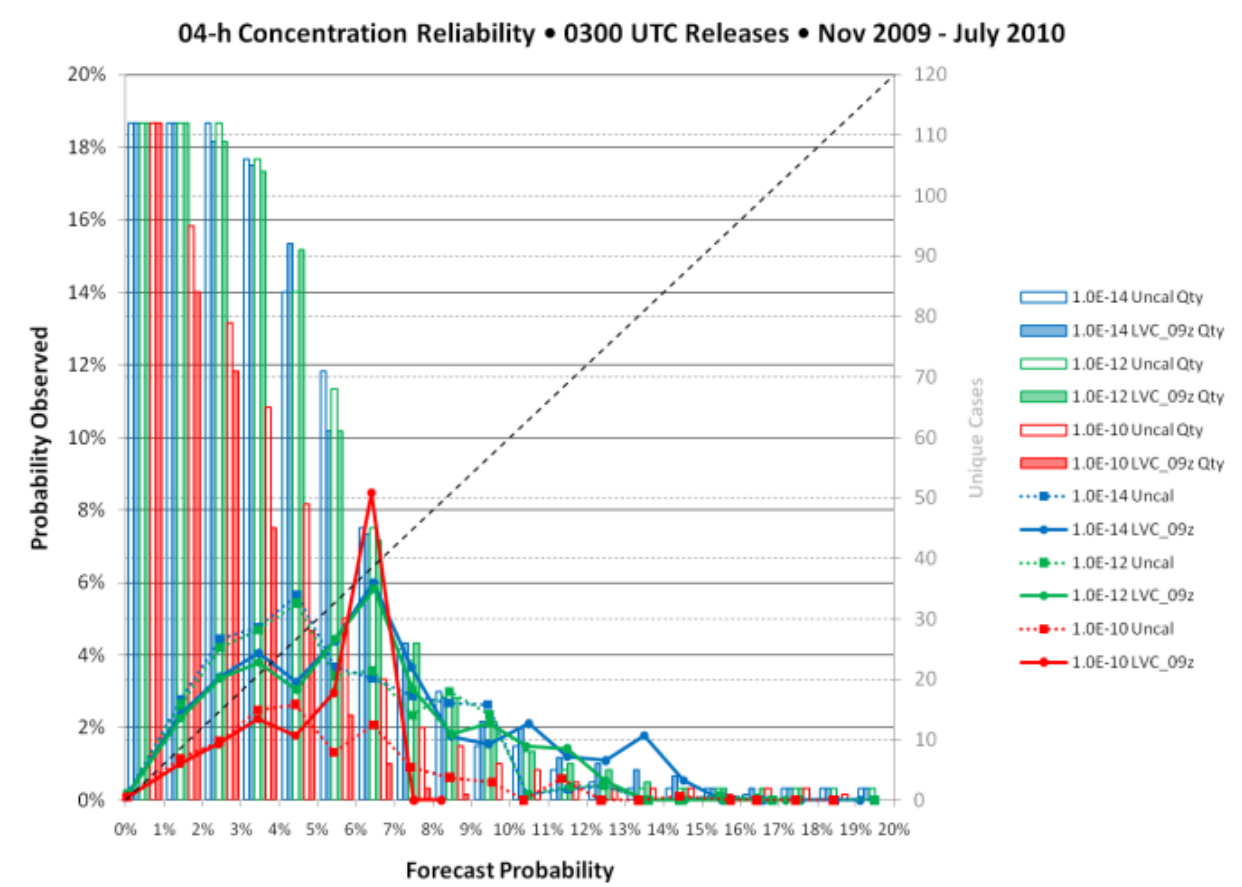

Figure 1 — Comparison of reliability of 04-h concentration forecasts for various thresholds using a fixed sampler grid with resolution of 0.05° × 0.05° divided into 1% bins for concentration thresholds of 10^{-14} kg·m^{-3} (blue), 10^{-12} kg·m^{-3} (green) and 10^{-10} kg·m^{-3} (red). Dotted lines indicate the reliability of the forecasts using uncalibrated wind variances. Solid lines indicate the reliability of the forecasts using calibrated wind variances. The dashed black line indicates perfect reliability. Bar graphs on secondary vertical axis (right) indicate the number of unique forecasts contributing to the probability bin, with unfilled bars corresponding to uncalibrated forecasts and filled bars corresponding to calibrated forecasts.

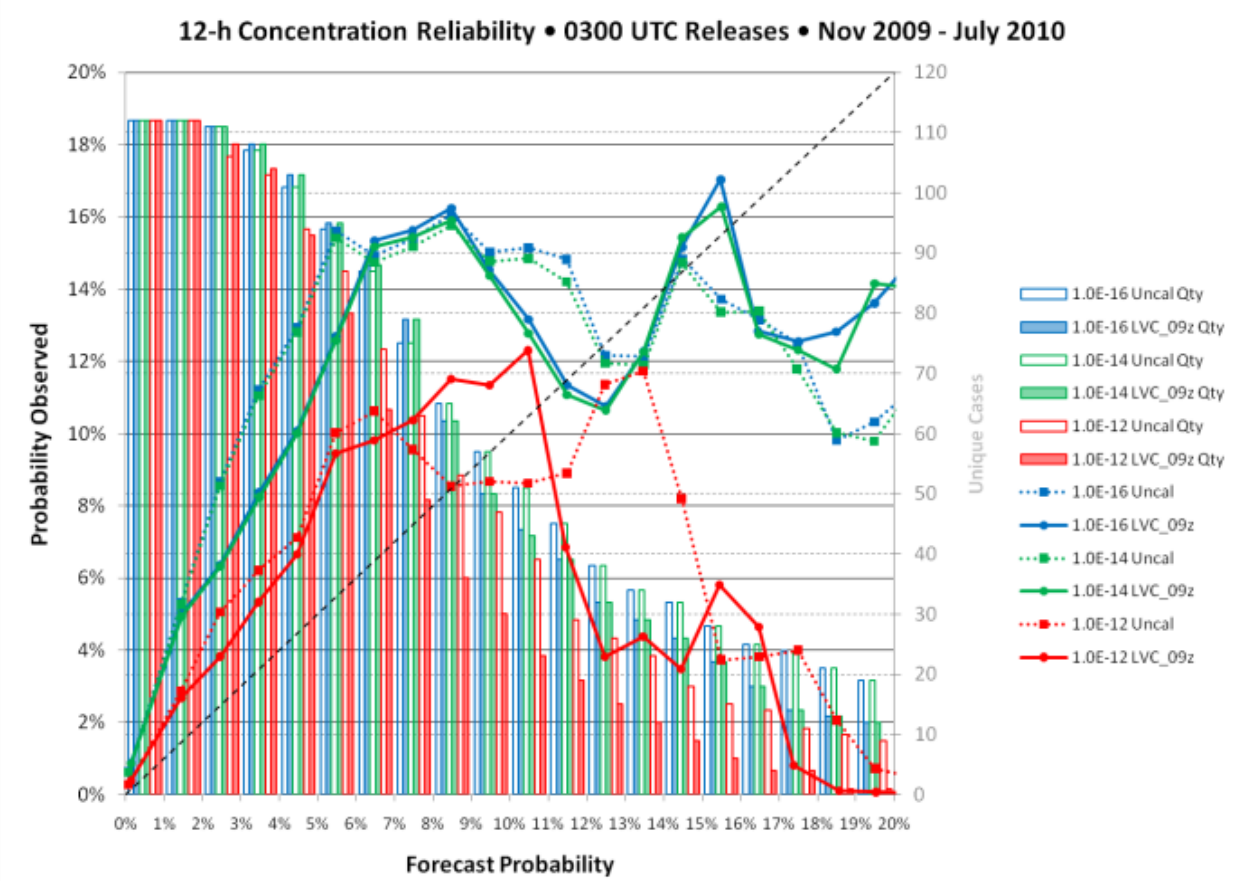

Figure 2 — As in Fig. 1, but at 12 h using concentration thresholds of 10^{-16} kg·m^{-3} (blue), 10^{-14} kg·m^{-3} (green), and 10^{-12} kg·m^{-3} (red).

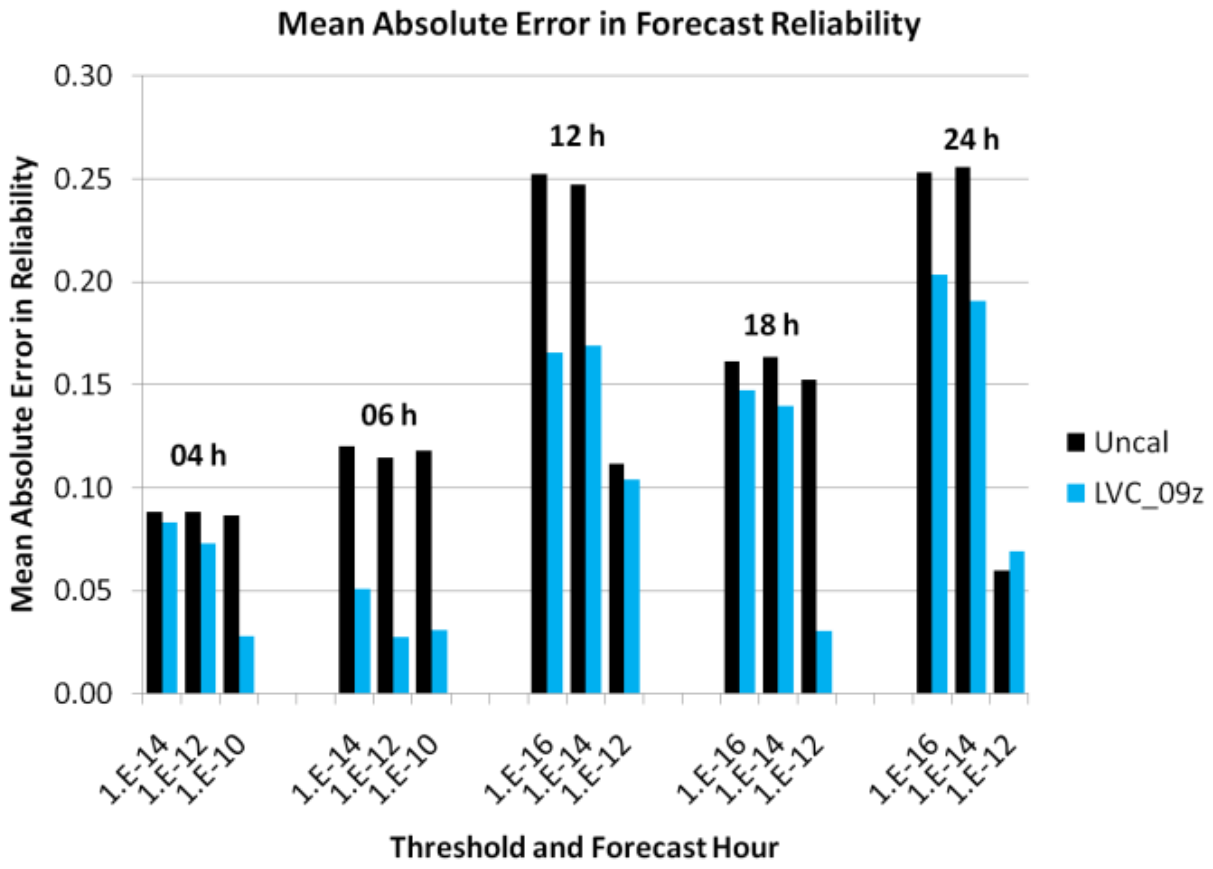

Figure 3 — Mean absolute reliability error (MARE) in the forecast reliability of forecasts at five different forecast lead times using three concentration thresholds (as indicated along the x-axis in kg·m-3). Black bars indicate the MARE of the uncalibrated forecasts; teal bars denote the MARE of the calibrated forecast.

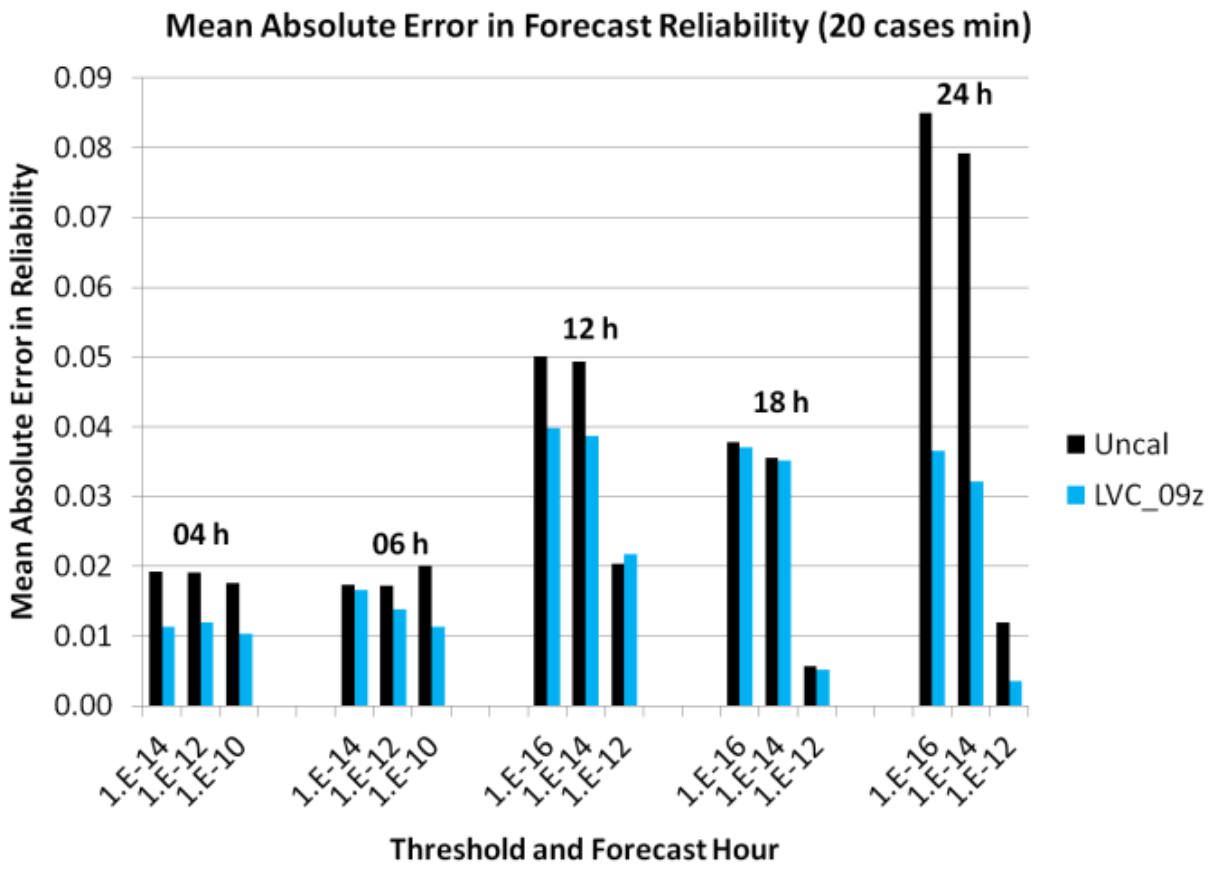

Figure 4 — As in Fig. 3, but computed using only bins with at least twenty different forecasts contributing.

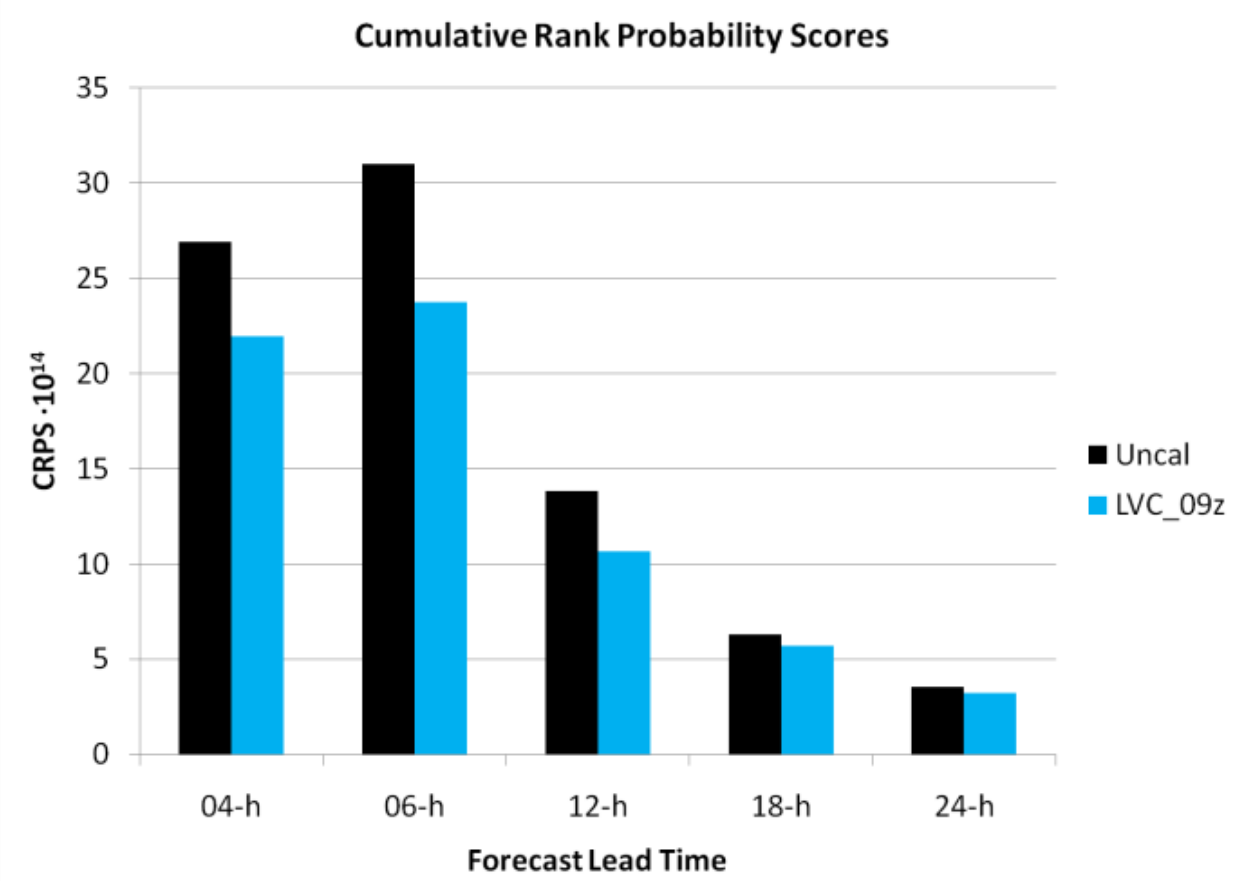

Figure 5 — Cumulative rank probability score (CRPS) of concentration forecasts at five different forecast lead times. Black bars indicate CRPS for forecasts using uncalibrated wind variances; teal bars indicate CRPS for forecasts using calibrated wind variances.

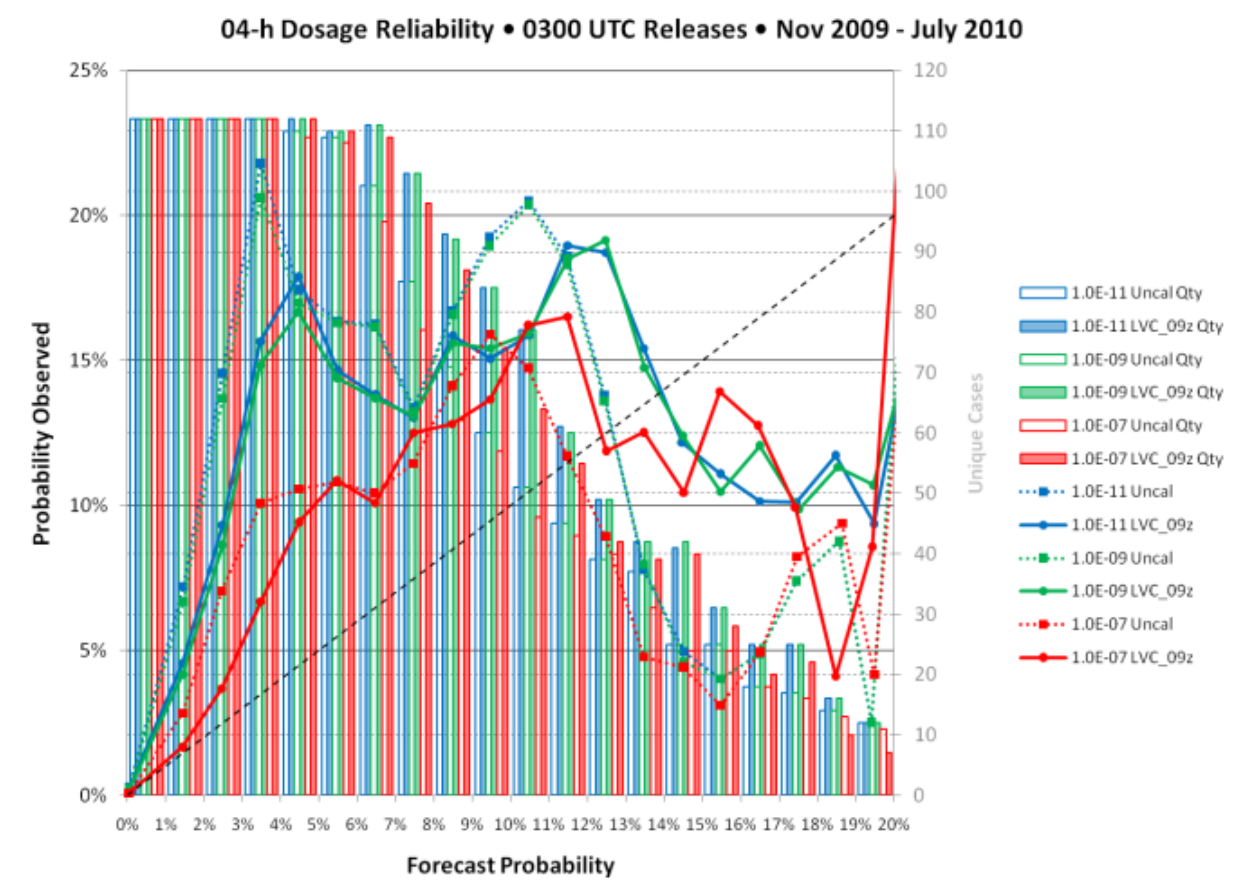

Figure 6 — As in Fig. 1, except for 4-h surface dosage at thresholds of 10^{-11} kg·s·m^{-3}(blue), 10^{-9} kg·s ·m^{-3} (green) and 10^{-7} kg·m^{-3} (red).

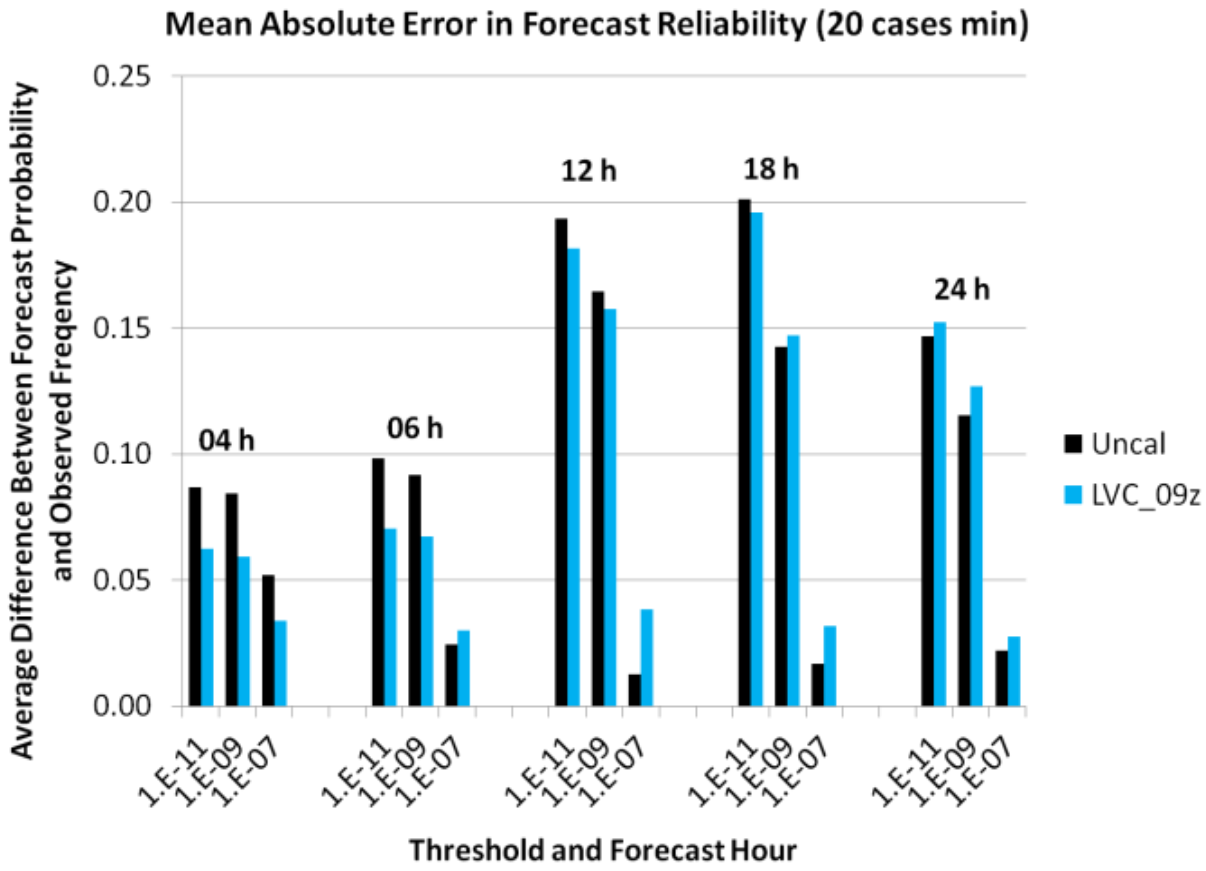

Figure 7 — As in Fig. 4, but for surface dosage forecasts. Thresholds along the x-axis have units of kg·s·m^{-3}.

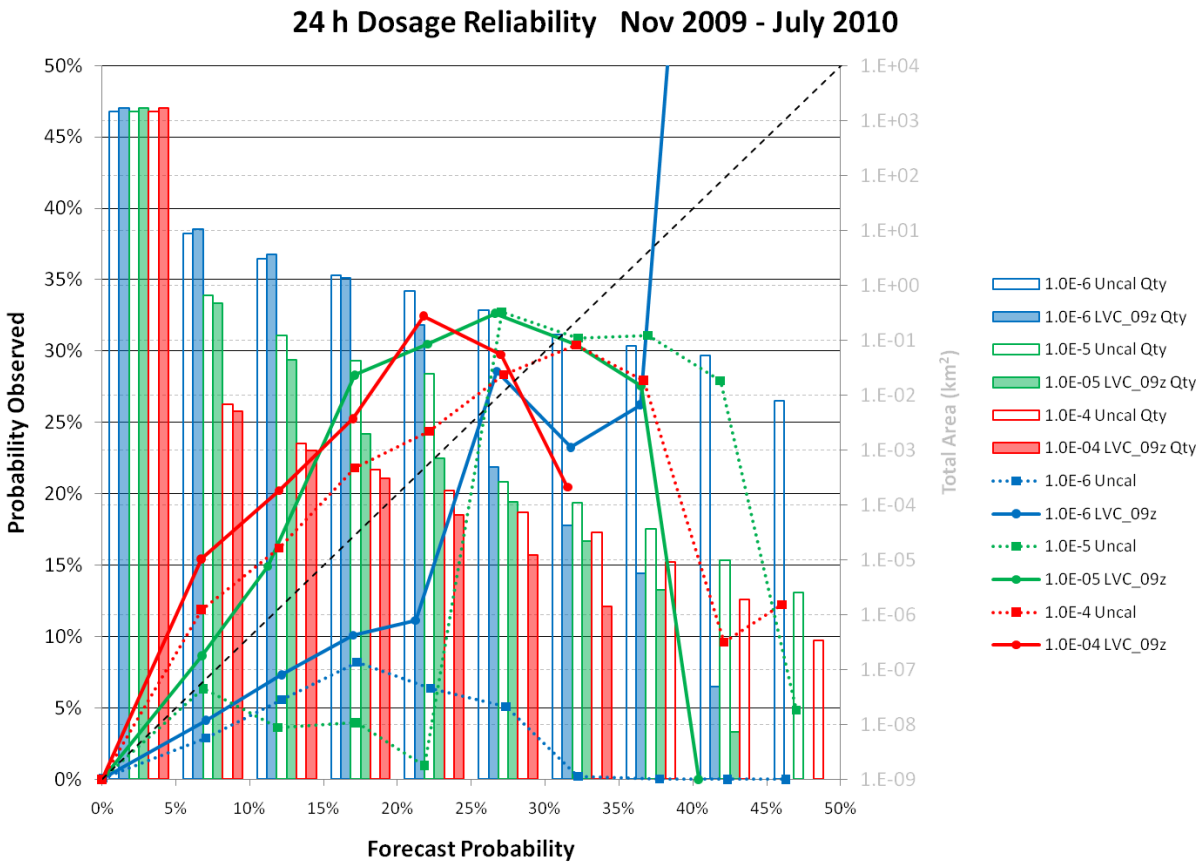

Figure 8 — Comparison of reliability of 24-h surface dosage forecasts for various thresholds using the SCIPUFF adaptive grid, weighted by area in 5% bins at surface dosages of 10^{-6} kg·s·m^{-3} (blue), 10^{-5} kg·s·m^{-3} (green), and 10^{-4} kg·s·m^{-3} (red). Dotted lines indicate the reliability of the forecasts using uncalibrated wind variances. Solid lines indicate the reliability of the forecasts using calibrated wind variances. The dashed black line indicates perfect reliability. Bar graphs on secondary vertical axis (right) indicate the total area in each forecast probability [log scale].