Verification and Calibration of Simulated Reflectivity Products

Mark T. Stoelinga

University of Washington

1. Introduction

Simulated reflectivity (hereafter, SR) is the equivalent radar reflectivity field that is calculated from the precipitation output of a numerical weather prediction (NWP) model. While research studies have made use of SR for examining model output for many years, such displays are becoming increasingly popular as a means for displaying forecast fields from high-resolution operational NWP models. In particular, plan view displays of composite SR (the maximum SR in a vertical column) have become a common post-processing product of both research and operational models. SR offers major advantages over traditional precipitation forecast displays, including the obvious fact that SR is easier to verify in real time by direct comparison with readily available, observed composite equivalent reflectivity products (Koch et al. 2005). It is also little wonder that a forecast field of radar reflectivity would be embraced in the severe weather and tropical cyclone forecasting communities, where the signatures of storm structures, evolutions, and motions are more recognizable in radar reflectivity fields than perhaps any other observed quantity. Another subtler advantage of SR is that it allows one to more easily see detailed mesoscale and storm-scale structures capable of being forecast by finer resolution NWP models. Examples demonstrating this advantage for a variety of mesoscale phenomena observed during recent forecast experiments in the continental United States were presented by Koch et al. (2005).

However, before any meaning can be ascribed to the SR products generated by NWP models, it is important to understand how they are calculated, and how they behave relative to observed reflectivity. First, SR is beholden to the fidelity of the model cloud and precipitation microphysics forecast, since it is derived directly from the hydrometeor mixing ratios. Any biases in those mixing ratios will be reflected in the SR field. Furthermore, a particular challenge in trying to produce a SR product (and trying to *reproduce* observed reflectivity) is the diameter-to-the-sixth-power dependence of equivalent reflectivity factor. This dependence renders reflectivity highly sensitive to the largest precipitation particles present, and thus renders SR highly sensitive not only to the precipitation mixing ratios, but to assumptions about the precipitation size distributions. It is conceivable that a model could be performing well in terms of precipitation forecast, but producing unrealistic SR fields due to poor representation of the particle size distributions. Considering these challenges in producing accurate SR fields, the purpose of the present study is to examine SR products from two different perspectives: First, we examine the behavior of existing reflectivity products compared to observations, in an attempt to ascertain where they are failing. This is accomplished by means of comparison of 3-D SR produced by two models with 3-D analysis mosaics of observed reflectivity for a single case study that occurred during the Developmental Testbed Center (DTC) Winter Forecast Experiment (DWFE, Bernardet et al. 2005). The fields are compared both by direct examination and through use of "contour frequency by altitude" (CFAD) diagrams (Yuter and Houze 1995). Second, we consider the following question: If it can be shown that there is a systematic error in a particular SR product, in terms of the frequency distribution of SR values compared to observed (i.e., if one model product consistently produces too much or too little echo of a given value), can the SR product be "calibrated" to more closely match the observed frequency distribution? We address this question first by examining the same single case study mentioned above, and show that a calibration is possible and, when applied, improves the "look" of the SR fields. We then examine an entire month of SR fields, and show that there is indeed a systematic error in the fields that is consistent with the experience of the forecasters that used them, and that this systematic error can be "calibrated away". This report concludes with a discussion about the merits of the calibration approach, caveats, and other potential uses.

2. The DTC Winter Forecast Experiment (DWFE)

During the winter months of early 2005, a focused effort to provide high-resolution NWP model guidance to forecasters was conducted by the Developmental Testbed Center (DTC). During this project, known as the DTC Winter Forecast Experiment (DWFE, Bernardet et al. 2005), several model-derived forecast products, including composite SR fields, were produced over the entire CONUS domain and made available for experimental forecasting purposes to the National Weather Service. The composite SR fields were produced from 5-km Weather Research and Forecasting (WRF) model forecasts made by the NCAR Advanced Research WRF (ARW) model and the NCEP Nonhydrostatic Mesoscale Model (NMM). As discussed in Koch et al. (2005), there were several differences in the development of the two SR products. Not only were two different dynamical models used, but each used a different microphysical scheme: the NMM model used the Ferrier et al. (2002) scheme (hereafter, “Ferrier”), whereas the ARW used the Hong et al. (1998) scheme (hereafter, “WSM5”). Both schemes are single-moment bulk schemes with four different classes of hydrometeors (cloud water, cloud ice, rain, and snow), and both assume exponential size distributions for the precipitation hydrometeors. However, the two schemes use different intercept parameters for the snow size distribution. To complicate matters further, the ARW SR product was calculated with an algorithm that assumed a snow intercept parameter different from that in the WSM5 scheme, whereas the NMM SR product was calculated in a manner internally consistent with assumptions in the Ferrier scheme. Koch et al. (2005) describe and quantify the dependence of the SR product on the snow intercept parameter used in the SR calculation. One of the purposes of the present study is to examine this effect in a DWFE case study.

3. Data

The SR products considered in this study are verified against two different national (conterminous United States) operational radar reflectivity “mosaic” products. The first is part of an experimental product under development at the National Severe Storms Laboratory, known at the National Mosaic and Multi-sensor Quantitative

Precipitation Estimation (NMQ). This mosaic analysis of the NWS WSR-88D operational radars is described in Zhang et al. (2004), and includes both a 3-D analysis and a 2-D composite product, each on a 1-km horizontal grid. The 3-D analysis is particularly useful for comparing with 3-D model SR, to help understand where the model's SR is performing well and where it is not. However, due to the tremendous size and format of the archived NMQ data set, it is somewhat unwieldy for use in verification of SR over a wide area and for a long period of time. For this purpose, 2-D raster images of both observed and modeled reflectivity were used. Using appropriate software, pixels of different colors, corresponding to different reflectivity values, could be counted over wide areas and for a large number of time periods. The observed composite reflectivity analyses originated from the WSI Corporation, and were a routinely available product during DWFE.

4. The 13 February 2005 stratiform precipitation event

Because the primary purpose of DWFE was the forecasting of winter weather rather than spring-time severe convection, a case study was chosen that included a large winter cyclonic storm system with significant areas of stratiform precipitation on the north and northwestern sides of the storm (Fig. 1). Also, we focus on a particular area of the stratiform precipitation shield in the upper Midwest, not for any meteorological reason, but simply because this is the area covered by one of the "tiles" of the archived NMQ radar mosaic data. The 1-km NMQ data and the 5-km NMM model output were analyzed to a subdomain of the 5-km ARW model grid covering the same region as the NMQ "tile", so that all three data sets could be compared on a common grid structure.

The composite reflectivity field is zoomed in to the "tile" subdomain in Fig. 2a. The other panels of Fig. 2 depict the SR products derived three different ways. Figure 2b is the SR product from the ARW model that was produced in real time. This product assumed a constant intercept parameter for the snow size distribution, even though that is not what is assumed in the WSM5 scheme. The WSM5 scheme assumes a temperature-dependent intercept parameter, which effectively skews the distribution toward larger particles at warmer temperatures, and toward smaller particles at colder temperature. To examine the effects of the inconsistent assumptions (between the model and the SR algorithm), the SR was recalculated with a model-consistent intercept parameter. In addition, an attempt to better capture the “brightband” was employed, in which snow at above-freezing temperatures was assumed to have the dielectric constant of liquid water. The result of this “corrected” SR product is shown in Fig. 2d. Finally, the lower left panel (Fig. 2c) shows the SR product from the NMM model, which is the same product that was produced operationally, and is fully consistent with the assumptions in the Ferrier scheme. Several differences between these products and observations are readily apparent. The NMM SR product (Fig. 2c) simultaneously produces an over-extension of very low echo over a wide area to the north, and an underestimate of maximum echoes in the strongest part of the rainband. The ARW product calculated with constant-intercept is similar to NMM in its underestimate of maximum echoes, but does not over-extend the low-echo region as NMM does. The ARW product calculated with model-consistent intercept also does not overextend the low-echo region, but significantly over-estimates the areal coverage of highest reflectivity values.

To help understand some of the behaviors seen in the composite SR displays in Fig. 2, CFADs (contour frequency by altitude diagram, Yuter and Houze 1995) were constructed for both the 3-D SR products and the 3-D NMQ analysis of observed radar reflectivity for this case. The CFADs utilize all grid points and all available vertical levels in the domain shown in Fig. 2. Calculation of the CFADs follows the standard procedure, with the following two exceptions: (1) Heights are referenced by height above (or below) the freezing level, rather than above ground level or MSL. This was done because a wide area was being considered, over which the freezing level height varies, but it is desirable to composite with reference to the freezing level, since that is typically an important transition level in the vertical profile of reflectivity. (2) In order to make quantitative comparisons of one model product to another or to observed, the usual normalization of the frequencies at a particular height by the total frequency at that height was intentionally omitted. Contour values shown have no meaning in an absolute sense, but relative differences between contour values at one height/reflectivity point and another in the same plot or in a different plot are quantitatively meaningful.

The CFADs for the observations and three different SR products are shown in Fig. 3. In a gross sense, they are all typical of stratiform precipitation: reflectivity increases downward to a maximum at the freezing level, below which it decreases slightly and then remains constant down to the surface (roughly 2 km below the freezing level). However, important differences are seen. The observed CFAD (Fig. 3a) shows reflectivity values decreasing below the brightband much more than any of the models. This is most likely due to a an artifact of the observational system, in which spatial coverage decreases as the vertical level decreases beneath the lowest scan cones of the NEXRAD network, artificially decreasing the frequencies of all dBZ values at those low levels. Bearing this in mind, the WRF-ARW with constant-intercept SR (used in real-time) actually appears to be the best match among the model products. The "corrected" ARW product enhances the brightband significantly, due to both the shift of the snow size distribution to larger particles at warm temperatures, and the inclusion of the wet snow dielectric factor correction, which increases reflectivity by up to 7 dB locally. The net result in the composite SR field (Fig. 2d) is a brightband-dominated field that overestimates the observed composite reflectivity by 5-10 dB.

The NMM CFAD (Fig. 3c) is the biggest outlier, and here is where the non-normalized contours really show how different NMM is behaving relative to ARW and observations. At upper levels NMM is producing very light echo (due to light snow) with very high frequency, starting at about 3 km above the freezing level (5-10 DBZ) and continuing upward. Frequency of occurrence of these low dBZ values aloft are ~5-10 times that in the observations or other SR products. If this light snow occurred only above larger SR values at lower altitudes, it would not effect the composite SR field. However, at many locations the light snow aloft produced the maximum SR in the column, thereby contaminating the composite SR field with echoes that bear little relevance to surface precipitation.

As indicated in the introduction, the SR products in this test case study were used to explore the idea of developing a mapping or calibration function that could be applied to the composite SR values, such that the composite SR field would resemble the observed in terms of the dBZ frequency distribution. Mathematically, this can be expressed as seeking a function $Z_{\mathrm{new}} = h(Z_{\mathrm{m}})$, such that

$$f(Z_\mathrm{m}) = g(h(Z_\mathrm{m}))\frac{dh}{dZ_\mathrm{m}} \quad ,$$

where Z_m is the composite SR, and $f(Z)$ and $g(Z)$ are the frequency distributions of the simulated and observed composite reflectivity, respectively. While $h(Z_\mathrm{m})$ is difficult to extract mathematically, there is a practical and simple way to arrive at it. Starting with a set of SR values that will be used to obtain the calibration equation (e.g., all the grid values of SR in a single plot such as Fig. 2a), first, all the values are ranked in order from lowest to highest value. Then the same is done for the corresponding observed reflectivity set. It is important that the same number of points is used for both. By aligning the two ranked sets (simulated and observed), the full set of pairs of reflectivity values provide the precise calibration function needed to transform the SR plot into one that has the exact same frequency distribution as the corresponding observed reflectivity plot.

This method was applied to the three SR fields in Fig. 2b-d, using the observed field in Fig. 2a. The resulting calibration curves are shown in Fig. 4. The first thing to note about these curves is that, although they deviate significantly from the one-to-one line, they are fairly linear themselves. It can be shown that if the frequency distributions of two variables are of the same functional form, even if they have different means and variances, then the two variables must be linearly related. Like many geophysical phenomena, observed radar reflectivity fields tend to have a log-normal distribution, i.e., its logarithm (or the dBZ value in the case of reflectivity) is normally distributed. This has been confirmed for the reflectivity field seen in Fig. 2a. Therefore, over the range that the calibration function is linear in Fig. 4, the SR field is also log-normally distributed, although not with the same log-normal distribution. In ranges of reflectivity where the calibration function is not linear, the SR field deviates substantially from a log-normal distribution, indicating a significantly “non-natural” behavior, such as at low reflectivities (SR < ~15 dBZ).

Another interesting feature of the calibration curves in Fig. 4 is that they almost entirely fall below the one-to-one line, meaning that in nearly all cases, the SR values need to be reduced in order to achieve the observed frequency distribution. At the light-precipitation levels (SR = 15 dBZ), the SR values need to be reduced by up to 20 dBZ.

The exception is for the higher reflectivities in the case of the NMM and ARW/constant-intercept products, which require an increase of reflectivity values by about 5 dBZ to achieve the observed frequency distribution. When the calibration curves are actually applied to the corresponding SR products (Fig. 5), the result is a reflectivity distribution that looks much more like the observed in terms of areal coverage of the different reflectivity color bands (compare Figs. 5 and 2).

Of course, by looking at only a single time period, the calibration curves arrived at here could reflect both a poor forecast in this one instance, and some systematic bias in the model and/or SR algorithm. Naturally, the systematic bias is of much greater interest for the purpose of verifying and/or calibrating the SR product. It would not make any sense to apply these calibration curves in an attempt to obtain a statistically more accurate SR product at all times and in all locations, since they could arise entirely from a poor precipitation forecast in this one instance.

To address this issue, a similar analysis of a four-week data set of observed and simulated composite reflectivity products was conducted, described in the next section.

5. Four-week analysis of composite reflectivity

Using composite radar imagery of both observed and SR products that were archived during a particularly active period of DWFE from 28 February – 24 March 2005, a set of calibration curves were obtained for the ARW/constant-intercept and NMM products. Three time periods were used per day (18, 21, and 00 UTC), and the region that was used covered most of the conterminous United States east of the Rocky Mountains. It should be noted that the inclusion of the southern U.S. in March, as well as using only afternoon time periods, resulted in a significant number of convective events in addition to the more stratiform winter precipitation that prevailed in the northern part of the domain. The calibration curves are shown in Fig. 6. Also shown are the calibration curves obtained for the same SR products for the single case study (i.e., the same curves depicted in Fig. 4). The long-term calibration curves are much closer to the 1-to-1 line in the range of SR = 10—30 dBZ, indicating that the behavior for the single case study was exaggerated due to an imperfect forecast at that particular time. However, the long-term curves still indicate that a significant reduction of light-precipitation echo is required to match the observed frequencies of these reflectivity values. The long-term curves cover a higher range of reflectivity values, due to the presence of convective echoes in the long-term data set. The NMM calibration curve reflects the general experience of forecasters during both DWFE and during the NSSL Spring Forecast Experiment in 2005, namely, that in addition to producing too much light precipitation echo, the NMM SR product almost never produced echo > 50 dBZ. Thus, the NMM long-term calibration indicates that 50-dBZ values should be increased to ~65 dBZ in order to match observed frequency distributions. The behavior of the ARW product's calibration curve was similar to, but less pronounced than, that of the NMM product. Thus, it appears that sufficiently pervasive systematic biases exist in the SR products, such that the use of calibration curves could result in more accurate results in terms of matching the observed frequency distribution better.

6. Caveats and merits of the use of calibration of SR products

While the calibration approach would seem to potentially provide improved SR products, several caveats must be considered. Calibration will not significantly improve correlations between observed reflectivity and SR. In fact, if the calibration is linear, it will not improve linear correlations at all, by definition. However, in assessing the value of a precipitation-related forecast product, it is often regarded as more important to have a realistic field that matches the character of the observed field, rather than one that is well-correlated with the observed field, a philosophy which has led to recent keen interest in object-oriented verification techniques (Ebert and McBride 2000). The calibration approach is in line with this philosophy.

The calibration curve for a particular SR product should obviously be based on a large data set, to ensure that it is not heavily influenced by a small number of poorly forecast systems. It should also be recognized that the calibration curve is likely to be dependent on a great number of factors associated with both the model and observations, including observational data quality, compositing method, model resolution and physics, SR algorithm, geographic location and time of year, etc. Thus, the use of calibration would be most appropriate for a fixed operational model configuration for which long-term statistics can be gathered. A single calibration equation would only be useful in a more general application in the case of very strong biases that occur ubiquitously, such as characteristic biases associated with a particular microphysical scheme. Finally, the fairly simple calibration approach presented here can certainly not recover from flaws in either model physics or SR algorithm. Note that when each of the two ARW SR algorithms were individually calibrated for the single case study, the solutions did not become identical (Figs. 5b and d).

One possible additional application of the calibration method is to provide a more reasonable forward-operator for variational methods that assimilate radar reflectivity data into high-resolution models. If the model has a systematic inability to produce precipitation hydrometeor fields similar to those in nature, it may be detrimental to try to force the observed reflectivity data into the model using a forward operator with which the model is not compatible. Using the calibration curves for a forward operator may provide a more reasonable connection between the model and observations.

7. Conclusions

This study has examined the behavior of simulated reflectivity (SR) products that are becoming increasingly popular among the forecasting community. Comparison of three different composite SR products that were routinely produced during the DTC Winter Forecast Experiment showed marked differences in their depictions of reflectivity compared both to each other and to an observed composite reflectivity data set. In one particular case study, contour frequency by altitude (CFAD) diagrams were used to elucidate the vertical differences in reflectivity that influenced the differences in composite reflectivity. It was found that even with the same model output (the WRF-ARW model run), differences in the assumed size distribution used in the SR algorithm resulted in significant differences in composite SR. Also noteworthy was the ubiquitous overprediction of areas of low echo by the NMM model, which resulted from excessive production over wide areas of ice aloft that never reached the ground.

In addition to examining the behavior of the SR products from the CFAD perspective, this study attempted to develop calibration functions which, when applied to the SR field, would result in a reflectivity field with a frequency distribution that matched that observed. For the single case study, calibration functions were found which reflected the errors in the different SR fields, and which, when applied to the SR fields, resulted in new SR fields that "looked by eye" to much more closely match the observed reflectivity field. A more comprehensive study was then conducted based on a larger area and longer-term data set, to ascertain if significant systematic long-term biases in the SR frequency distributions existed, and could be removed with calibration. Although the calibration curves obtained were not as far from the 1-to1 line as they were in the single case study, they did deviate by more than 10 dBZ at some points along the reflectivity spectrum. Various caveats and potential benefits of using calibration for simulated reflectivity were discussed.

Acknowledgements. The author wishes to thank the NSF and the staff of DTC for the rewarding opportunity to participate in the DTC Visitor Program. This is a valuable program that should be not only continued, but expanded.

References

Bernardet, L.R., L.B. Nance, H.-Y. Chuang, A. Loughe, M. Demirtas, S. Koch, and R. Gall, 2005: The Developmental Testbed Center Winter Forecasting Experiment. Preprints, *21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction*, Washington, D.C., Amer. Meteor. Soc., CD-ROM.

Ebert, E. E., and J. L. McBride, 2000: Verification of precipitation in weather systems: determination of systematic errors. *J. Hydrology*, **239**, 179-202.

Ferrier, B. S., Y. Jin, Y. Lin, T. Black, E. Rogers, and G. DiMego, 2002: Implementation of a new grid-scale cloud and precipitation scheme in the NCEP Eta model. Preprints, *15th Conf. on Numerical Weather Prediction,* San Antonio, TX, Amer. Meteor. Soc., 280-283.

Hong, S.-Y., H.-M. H. Huang, Q. Zhao, J. Dudhia, and S.-H. Chen, 2004: A revised approach to ice microphysical processes for the bulk parameterization of clouds and precipitation. *Mon. Wea. Rev.*, **132**, 103–120.

Koch, S. E., B. Ferrier, M. T. Stoelinga, E. Szoke, S. J. Weiss, and J. S. Kain, 2005: The use of simulated reflectivity fields in the diagnosis of mesoscale phenomena from high-resolution WRF model forecasts. Preprints, *11th Conference on Mesoscale Processes / 32nd Conference on Radar Meteorology*, Albuquerque, New Mexico, Amer. Meteor. Soc., CD-ROM.

Yuter, S. E., and R. A. Houze Jr., 1995: Three-dimensional kinematic and microphysical evolution of Florida cumulonimbus. Part II: Frequency distributions of vertical velocity, reflectivity, and differential reflectivity. *Mon. Wea. Rev.*, **123**, 1941–1963.

Zhang, J., K. Howard, W. Xia, C. Langston, S. Wang, and Y. Qin, 2004: 3.5 Three-dimensional high-resolution national radar mosaic. Preprints, *11th Conference on Aviation, Range, and Aerospace*, Hyannis, Massachusetts, Amer. Meteor. Soc., on CD-ROM.

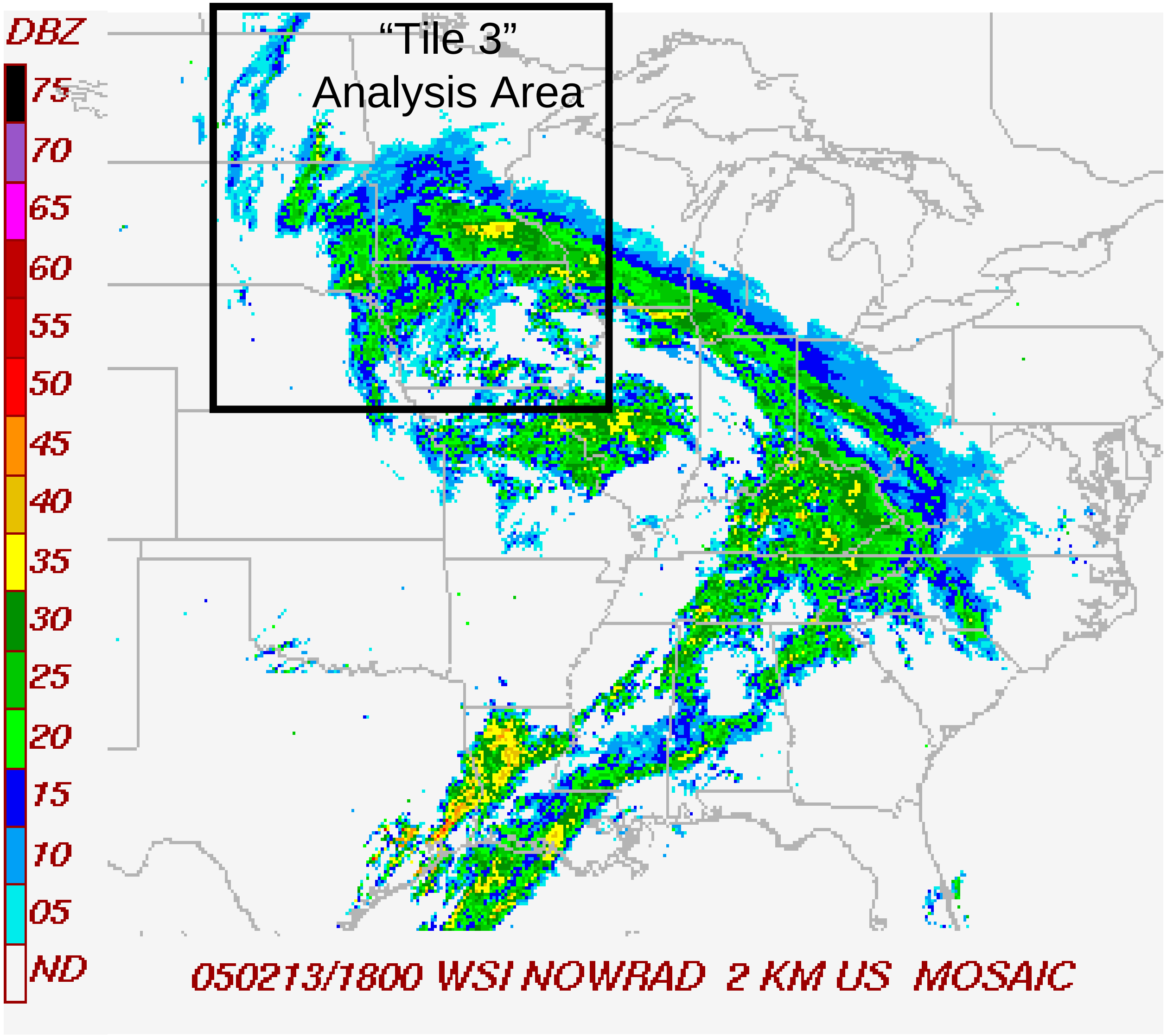

Figure 1. Composite reflectivity mosaic for 1800 UTC 13 February 2005. The box indicates the NMQ "tile 3" analysis area that is the focus of the case study.

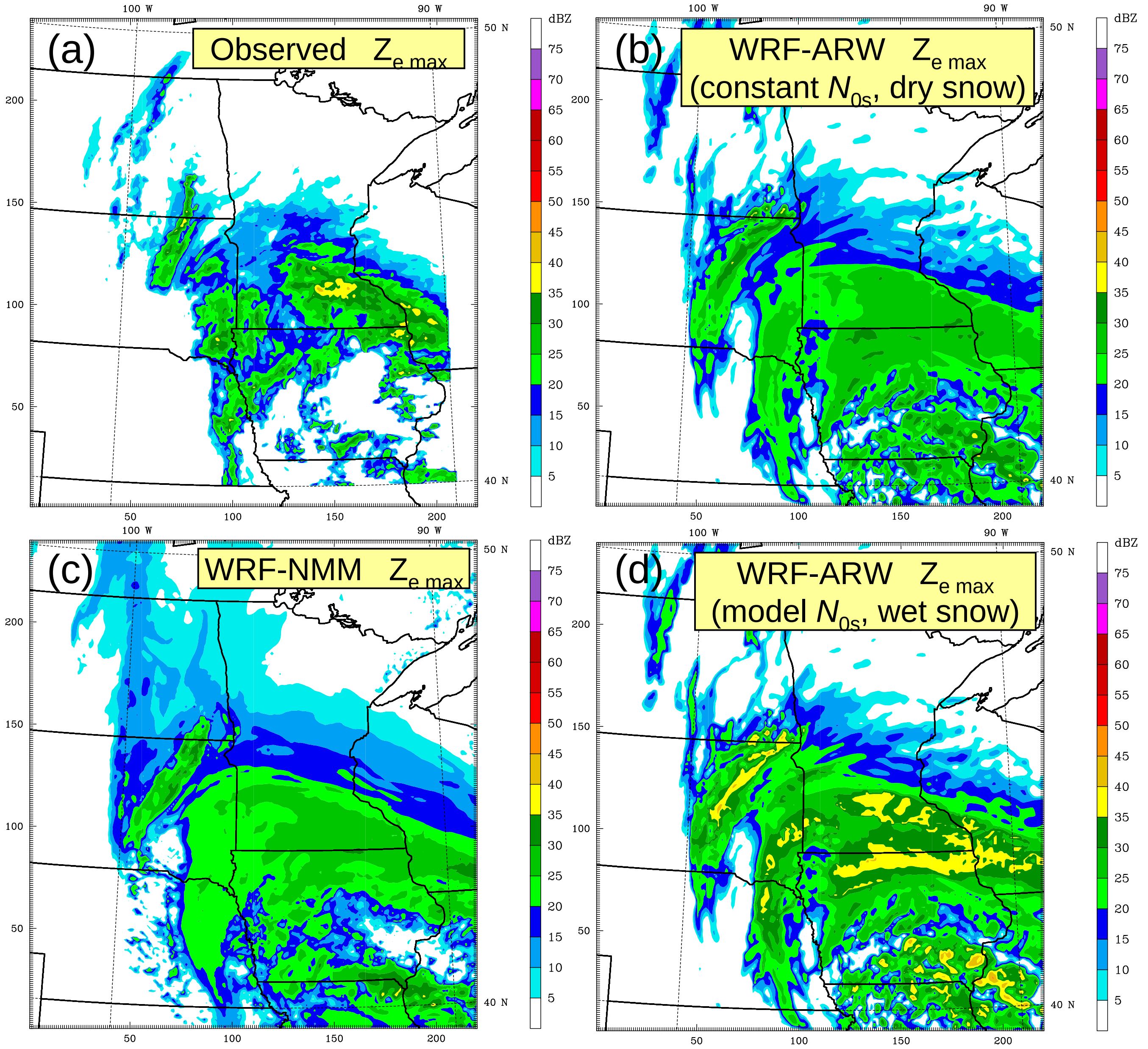

Figure 2. Composite radar reflectivity products over the “tile 3” area valid 1800 UTC 13 February 2005. (a) NMQ observed analysis; (b) simulated reflectivity, using constant-intercept and dry snow assumptions, from 18-h WRF-ARW forecast; (c) simulated reflectivity, using variable intercept and wet snow assumptions, from 18-h WRF-ARW forecast; (d) simulated reflectivity from 18-h WRF-NMM forecast.

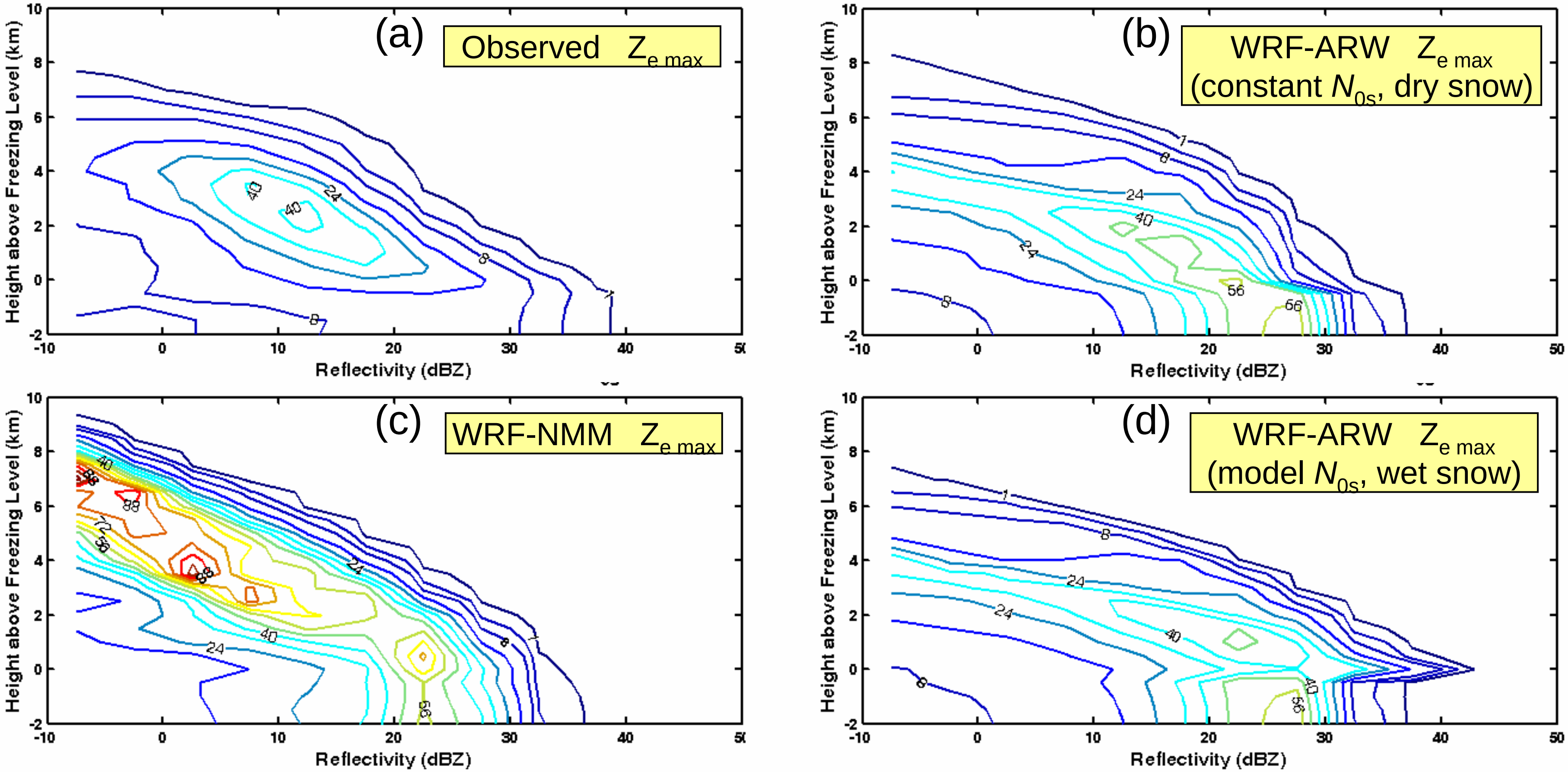

Figure 3. CFADs from 3-D reflectivity calculated in the “tile 3” analysis area. Panels (a)-(d) correspond to the same panels in Fig. 2.

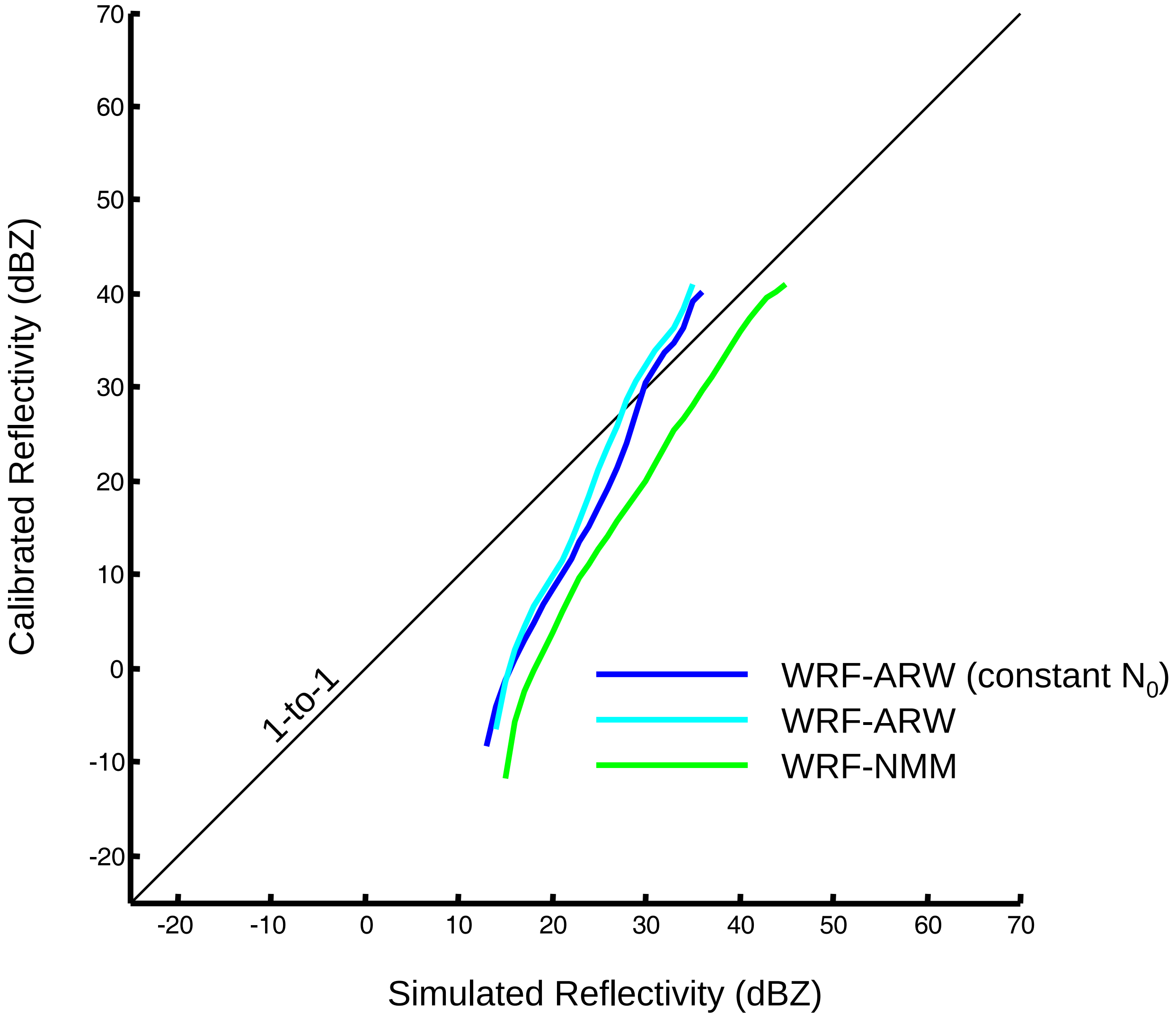

Figure 4. Calibration curves for composite simulated reflectivity fields shown in Fig. 2

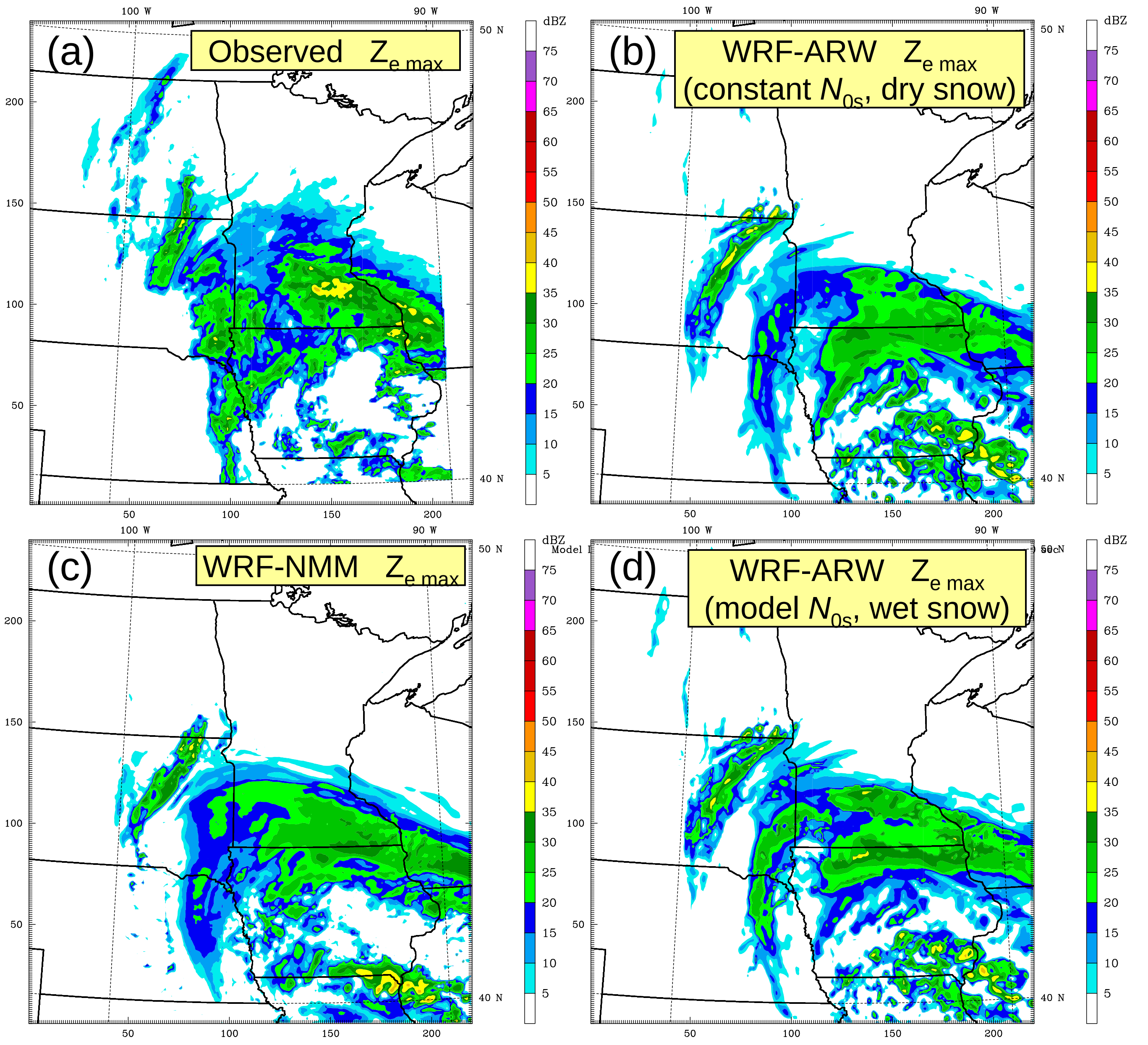

Figure 5. As in Fig. 2, but simulated reflectivty products have been calibrated to match frequency distribution of observed reflectivity field (panel a).

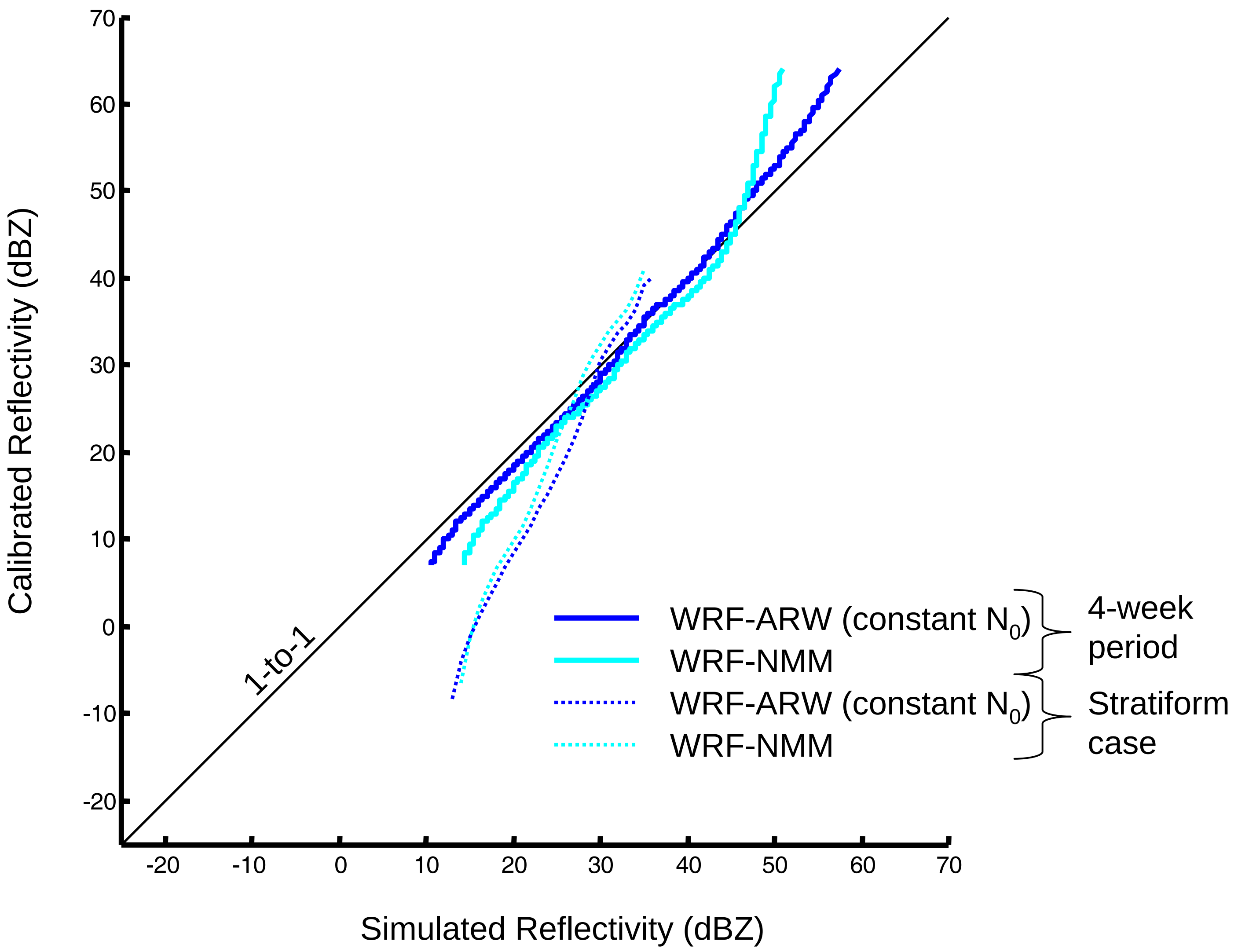

Figure 6. Calibration curves for composite simulated reflectivity fields throughout the ~4-week period of 28 February – 24 March 2005. Also shown for comparison are the calibration curves for the single case study shown in Fig. 4.